

Experimental evaluation of multimodal machine learning models for cardiovascular conditions in ICU patients



Felix Krones, M.Sc.

Wolfson College
University of Oxford

Supervisor: Prof. Adam Mahdi

Thesis submitted in partial fulfilment of the requirements for the degree of DPhil in Social Data
Science at the Oxford Internet Institute, University of Oxford

Hilary Term 2026

Abstract

Intensive care units (ICUs) generate vast amounts of heterogeneous data from various sources, ranging from imaging and waveforms to tabular records. Whereas clinicians routinely synthesise these diverse sources of information to guide patient care, traditional automated clinical decision-support systems have rarely integrated data across multimodal channels. Machine learning has the potential to replicate this synthesis through multimodal approaches, though the conditions under which accuracy genuinely improves remain unclear.

However, the development of such systems has been constrained by a lack of accessible, linked datasets and a limited understanding of data fusion methodologies. Consequently, how these data modalities interact remains largely unexplored. Specifically, there is a lack of systematic understanding regarding the specific conditions under which multimodal machine learning improves predictive performance, rather than simply introducing redundancy or algorithmic shortcuts.

To address this gap, this thesis establishes a framework to evaluate the performance of different methods for combining information from heterogeneous data sources, including imaging (X-rays), waveforms (ECGs), text (clinical notes) and tabular data (lab tests). The primary objective is to determine whether such integration improves model accuracy in an ICU setting and under which specific conditions it improves performance. Answering these questions became possible only recently following the accessibility of the MIMIC-IV database in 2021 and its linkage to ECG waveforms in 2023. Using cardiovascular conditions, a leading cause of mortality, as a critical test case, the thesis addresses, for the first time at large scale, challenges along the entire data and modelling pipeline. These span from data acquisition and processing, which I address by including a new methodology for digitising paper-based records, to the integration of waveforms within multimodal pipelines and the final combination of multimodal inputs.

The results show that multimodal integration improves predictive accuracy only conditionally. Late fusion proves most robust under real-world data constraints, with the largest gains in high-uncertainty cases and within the first 24 hours of admission. Subgroup analysis reveals that per-group calibration helps only under specific distributional conditions and that fairness constraints selected on validation data do not reliably transfer to test data. These findings shift the focus from whether multimodal fusion helps to when, how and for whom it does so, offering a principled framework for evaluating multimodal prediction systems.

Key words: data fusion, multimodal machine learning, medical imaging, clinical decision-making

Approx. 66k words

(Based on Microsoft Word's word count, without references.)

Contents

1	Introduction	5
1.1	What is this thesis about?	5
1.2	Background, gaps and research questions	6
1.3	Thesis structure and chapter contributions	11
1.4	Publications	14
2	Review of multimodal machine learning approaches in healthcare	17
2.1	Introduction	18
2.2	Data modalities	21
2.3	Model development	31
2.4	Fusion approaches	38
2.5	Multimodal datasets and studies	45
2.6	Conclusions and future research	47
3	Digitising ECG waveforms for use in multimodal pipelines	54
3.1	Introduction	55
3.2	Related work	56
3.3	Methodology	66
3.4	Results	72
3.5	Discussion	73
4	A generalisable approach for multimodal waveform integration	77
4.1	Introduction	78
4.2	Related work	80
4.3	Methodology	83
4.4	Results	87
4.5	Discussion	90
5	Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders	94
5.1	Introduction	95

5.2	Methodology	100
5.3	Results	115
5.4	Discussion	125
6	Ablation on subgroup differences in multimodal data fusion	138
6.1	Introduction	139
6.2	Related work	140
6.3	Methodology	143
6.4	Results	149
6.5	Discussion	154
7	Discussion	157
7.1	Principal findings	157
7.2	Strengths and limitations in context	159
7.3	Implications	163
7.4	Future research	165
A	Extended literature	169
A.1	Technical concepts	169
A.2	AI deployment challenges	180
A.3	Reliability in medical computer vision	182
B	Lessons learned from multiple years contributing to PhysioNet	193
B.1	2022: Transfer learning for heart sound classification	194
B.2	2023: Model robustness for ECG classification	194
B.3	2024: Data digitisation for ECG classification	195
B.4	Overarching Challenge lessons	195
C	Supplementary material to Chapter 2	197
D	Supplementary material to Chapter 5	203
E	Supplementary material to Chapter 6	220

1 Introduction

1.1 What is this thesis about?

This doctoral thesis investigates the conditions under which multimodal machine learning improves the prediction of cardiovascular conditions in Intensive Care Unit (ICU) patients.

While multimodal machine learning has the potential to replicate how clinicians synthesise information in everyday practice, the interactions that occur among the diverse data modalities when fused through a machine learning model remain largely unexplored (see [Chapter 2](#)). Specifically, there is a lack of systematic evidence regarding the specific conditions under which multimodal integration genuinely improves predictive performance rather than simply introducing redundancy (see [Chapter 5](#) and [Chapter 6](#)).

This thesis tackles these challenges by making the following contributions: First, I address foundational gaps in multimodal pipelines by introducing novel methods to digitise paper-based Electrocardiograms (ECGs) ([Chapter 3](#)) and a generalisable, spectrogram-based approach to integrate high-frequency waveforms into deep learning models ([Chapter 4](#)). These methods build on two works from the prestigious PhysioNet machine learning competition. The first proposed the most effective method for reconstructing ECG signals from printed records, winning the 2024 PhysioNet Challenge internationally. The second received the Best Preprint Award at the 2023 Computing in Cardiology (CinC) Conference. Second, I introduce a systematic evaluation framework that compares early, intermediate and late data fusion strategies across six cardiovascular conditions ([Chapter 5](#)). This analysis moves beyond simple performance metrics to identify not just the best performing fusion strategy under ICU data constraints, but also the specific time windows and uncertainty thresholds where multimodal data adds the most clinical value. Third, I investigate calibration techniques for multimodal data fusion to reduce subgroup disparities ([Chapter 6](#)).

This interdisciplinary research addresses challenges across the entire machine learning pipeline, from the acquisition and digitisation of analogue data to the evaluation of fusion strategies and fairness interventions. Using only publicly available datasets, the proposed approaches are designed to be reproducible and to serve as a foundation for future validation across clinical environments.

Developed in close collaboration with clinicians, this work aims to move multimodal prediction from ad hoc data combinations toward a principled understanding of when and how integration improves diagnostic accuracy. The clinical value of this understanding lies less in average accuracy gains than in knowing *when* and *which* additional data change a prediction, so that clinicians can act within the critical early hours of an ICU stay and on the most uncertain cases. The clinical case for early detection is long established: simple early warning scores, used at the bedside for years, discriminate in-hospital death within 48 hours with an area under the curve of 0.88 to 0.93 (Smith et al., 2014), and a systematic review of their use reports a trend towards lower in-hospital mortality after their introduction, particularly when coupled with a rapid-response outreach service, although the evidence across studies remains mixed (Alam et al., 2014). More recently, machine learning based decision support has translated this principle into measurable gains when deployed prospectively in hospitals. A sepsis early warning system lowered in-hospital mortality by a relative 18.7% among patients whose alert was confirmed within three hours (Adams et al., 2022), and an automated deterioration alert across 19 hospitals reduced 30-day mortality with an adjusted relative risk of 0.84 (Escobar et al., 2020).

1.2 Background, gaps and research questions

1.2.1 Background and gaps: Opportunities and challenges in healthcare

Machine learning approaches have shown potential to improve clinical decision-making and address major challenges of the healthcare system (Pinsky et al., 2024). Such challenges include the constant pressure from rising costs (Keehan et al., 2020) and a

global shortage of healthcare providers (Kirch and Petelle, 2017), while the demand for more personalised medicine (Zhang et al., 2018) and better outcomes continues to rise (Topol, 2019). Moreover, while costs increase continuously (Keehan et al., 2020), about 25% of spending may potentially be lost due to preventable errors and inefficiencies in administrative systems (Kocher, 2021).

Machine learning approaches offer the potential to improve clinical decision-making, address mentioned healthcare challenges and improve operational efficiency (Rajpurkar et al., 2022). However, to date, only few machine learning applications have been successfully adopted in clinical practice (Wu et al., 2023). For instance, out of 11 billion Current-Procedural-Terminology code claims analysed from the US from 2018 to mid-2023, only about 0.001% involved Artificial Intelligence (AI) associated procedures. Additionally, it was found that about 95% of those claims were associated with just three conditions: coronary artery disease, coronary atherosclerosis and diabetic retinopathy (Wu et al., 2023). This is despite the fact that by 2023, over 600 AI technologies had already received FDA approval, a figure that increased to over 1,300 by mid 2025 (FDA, 2025). Today, machine learning-based systems that have been implemented in clinical practice are mainly unimodal, partially extended by some background information; examples include image-based skin cancer screening (FotoFinder Systems GmbH, 2026), diabetic retinopathy screening (Healthvisors, 2026), stroke and radiology triage systems (Viz.ai, 2026) and vital-sign-based early warning systems (Philips Healthcare, 2026).

The limited real-world adoption of machine learning in clinical practice underscores the importance of developing approaches that align well with clinical workflows. However, as clinicians routinely integrate multiple data sources and modalities to inform decisions (Acosta et al., 2022; Amal et al., 2022), the reliance on mostly single data modalities in current machine learning applications has created a disconnect with clinical practice. This is particularly a concern in ICUs, where clinicians draw on many different data sources to make rapid, high-stakes decisions; these sources range from radiological imaging (e.g. X-rays) and high-frequency waveforms (e.g. ECGs) to structured tabular data and unstructured clinical notes (Acosta et al., 2022; Amal et al., 2022). While ICUs

currently rely on various scoring systems that use pre-selected variables to generate simple risk metrics, for example for organ failure assessments (Vincent et al., 1996) and for predicting length of stay and mortality (Zimmerman et al., 2006), these tools are often applied as broad prognostic instruments rather than for condition-specific decision support. Crucially, they typically focus on a predetermined set of variables and a single data modality, largely ignoring the full spectrum of available clinical information. Due to the importance of synthesising heterogeneous data modalities in ICUs for rapid decision-making, the use of multimodal machine learning approaches in the ICU holds great promise (Amal et al., 2022).

However, there is an absence of a systematic approach to multimodal data fusion. This gap is primarily due to the lack of publicly available multimodal datasets, resulting in the development of predominantly single-modality models (Amal et al., 2022; Pinsky et al., 2024). In our recently published review (Krones et al., 2025), we showed that even among studies using multimodal data, most models combine only a few data modalities, often in an ad hoc manner, failing to replicate clinical workflows where clinicians integrate information from multiple sources. Only a few studies have compared different fusion approaches and these comparisons have been limited in scope with respect to the data sources involved and produced varying results, highlighting the need for a more structured methodology. For instance, Huang et al. (2020c) and Imrie et al. (2025) found that late fusion performed best for pulmonary embolism and dermatological image classification, respectively, while Guo et al. (2019), Holste et al. (2021) and Hayat et al. (2022) reported superior performance using either early or intermediate fusion techniques for tumour segmentation, breast cancer classification and mortality prediction. Moreover, the benefits of incorporating additional data modalities remain largely unexplored for the clinical diagnosis of cardiovascular conditions in ICU patients specifically. In particular, understanding which features or modalities contribute the most predictive power and under what conditions, a key clinical need, has received little attention. This is concerning, as collecting new medical data is expensive and different data types usually become available only at specific time points during a hospital stay.

These gaps motivate the research questions which are outlined in the following part.

1.2.2 Research questions and case studies

Our literature review (Krones et al., 2025) revealed a critical disconnect in current research: while multimodal data are abundant in daily ICU settings, limited availability of multimodal data for machine learning research has led to the under-exploration of systematic approaches for integrating different modalities in machine learning models for clinical decision making. Current approaches often rely on ad hoc combinations or expert-crafted features that do not scale, limiting their ability to inform clinical decision-making reliably. This thesis addresses this gap through the following overarching research question:

RQ: *Under which conditions do multimodal approaches improve the prediction of cardiovascular conditions in ICU patients?*

I investigate multimodal machine learning approaches by focusing on cardiovascular conditions as a case study, as such conditions show high preventability, yet are a leading cause of mortality: cardiovascular and cardiopulmonary diseases remain the leading cause of death globally, with cardiovascular diseases alone accounting for 17.9 million deaths, or 32% of the global total (World Health Organisation, 2021, 2023). Notably, up to 80% of these deaths could be prevented through early detection and risk management (Raleigh et al., 2022). These conditions represent a significant proportion of ICU admissions (Anandaciva, 2020; Cook et al., 2022), making them critical targets for advanced diagnostic methods.

I primarily use the MIMIC-IV database (Johnson et al., 2024), which reflects complex, multimodal clinical environments. Additionally, I examine specific foundational challenges using targeted datasets (e.g. PTB-XL (Wagner et al., 2020) and I-CARE (Amorim et al., 2023)) to establish the necessary capabilities for broader multimodal integration.

To answer the overarching question, the thesis first addresses the fundamental challenge

of data representation. Clinical waveforms, such as ECGs, are often stored in paper format or require domain-specific feature engineering that limits generalisability. To overcome these barriers and enable scalable ingestion of time-series data into deep learning pipelines, I introduce a new methodology to digitise paper-based waveforms, outperforming the benchmark by a factor of two, and contribute a generalisable deep learning approach to integrate waveforms into multimodal pipelines that does not rely on expert knowledge.

With robust representations established, the focus shifts to the integration methodology within the ICU environment. It remains unclear whether complex fusion strategies offer tangible benefits over simpler approaches, or how the timing of data availability impacts predictive power. Furthermore, the specific conditions under which multimodal integration adds genuine predictive value need to be quantified. This motivates the first subsidiary research question:

SQ 1: *How do different data fusion strategies compare in performance and under which conditions does multimodal integration outperform single-modality baselines?*

Finally, high predictive performance alone does not guarantee clinical safety or robustness. Multimodal models risk subgroup disparities and poor generalisability under distribution shifts (Seyyed-Kalantari et al., 2021; Yang et al., 2024). To ensure the robustness of the proposed frameworks, the thesis asks:

SQ 2: *How can subgroup disparities be identified and mitigated in a multimodal setup?*

The primary data source for the core fusion and fairness experiments is the MIMIC-IV database (Johnson et al., 2024). Its selection is driven by the recent release of linked modules (MIMIC-CXR (Johnson et al., 2019b), MIMIC-IV-ECG (Gow et al., 2023) and MIMIC-Note (Johnson et al., 2023a)), which for the first time allows for the large-scale linkage of imaging, high-frequency waveforms, clinical notes and tabular data for each ICU stay. This unique data environment enables the systematic investigation of the

research questions in a setting that closely reflects the complex, data-rich reality of modern intensive care.

In the literature ([Baltrušaitis et al., 2018](#); [Barua et al., 2023](#); [Bayoudh et al., 2021](#)), the term ‘multimodal’ has been used to describe three distinct concepts: multimodal inputs (e.g. text and images), multimodal outputs (e.g. images with text descriptions) and cross-modal tasks involving different input and output modalities (e.g. text-to-image generation). In this thesis I use the term specifically in reference to the first: predictive models based on multimodal inputs, such as the integration of tabular, imaging and time-series data.

Besides multimodal research, other recent advancements have played a key role in enabling AI applications in healthcare, including but not limited to transfer learning, self-supervised learning, Convolutional Neural Network (CNN) and transformer architectures. These methodologies have improved the performance and versatility of AI models, paving the way for more precise and efficient solutions to a range of healthcare issues. While my research focus is on multimodal methods, I make use of these methodologies throughout this thesis and describe them in Appendix A.

1.3 Thesis structure and chapter contributions

This integrated thesis is organised around five research papers that I led as either the sole or joint first author, working alongside experts in medical machine learning, imaging and algorithmic fairness ([Table 1](#)). To ensure clinical relevance, the research was conducted in close collaboration with ICU clinicians. Each chapter addresses a specific gap in the multimodal prediction pipeline for intensive care settings:

Chapter 2 (Review of multimodal machine learning approaches in healthcare) systematically surveys the landscape of multimodal machine learning in healthcare. It categorises existing data modalities, fusion strategies and model architectures, while synthesising evidence from recent literature to identify where current approaches fail to replicate clinical decision-making.

By pinpointing the lack of systematic evaluation frameworks for data fusion, this review establishes the motivation for the subsequent chapters. It highlights the critical disconnect between the availability of multimodal data and the scarcity of robust, clinically aligned fusion methodologies, framing the specific research gap this thesis intends to fill.

In Appendix A, I provide a wider literature review, focusing on technical concepts, such as fusion techniques, model architectures and reliability research.

Chapter 3 (Digitising ECG waveforms for use in multimodal pipelines)

introduces a novel computer vision pipeline for digitising paper-based ECG records. The proposed method, which combines Hough transforms with deep learning segmentation, was benchmarked during the PhysioNet Challenge 2024, where it outperformed the state-of-the-art benchmark by achieving twice the validation score of the second-best entry.

Despite the shift to digital records, paper-based signals remain common, particularly in historical archives and due to their ease of transfer and the ability to annotate them manually. By enabling the reliable digitisation of these records, this method unlocks vast amounts of previously inaccessible data for modern multimodal pipelines, improving the diversity and generalisability of training datasets available for clinical research.

Chapter 4 (A generalisable approach for multimodal waveform integration)

proposes a generalisable computational approach for integrating high-frequency time-series data (such as ECGs and EEGs) into multimodal deep learning models. It converts raw signals into spectrograms to leverage standard image-based architectures (CNNs), demonstrating performance comparable to expert-derived feature extraction techniques without requiring domain-specific manual engineering.

Traditional waveform analysis relies heavily on labour-intensive, expert-crafted features that often fail to generalise across tasks. This spectrogram-based approach offers a scalable, data-driven alternative, removing a significant bottleneck in the development of clinical decision-support systems that utilise waveform data.

Chapter 5 (Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders) systematically evaluates early, intermediate and late fusion strategies for predicting clinically relevant cardiovascular conditions in ICU patients. Through structured ablation studies, it identifies the specific conditions under which multimodal integration improves performance, specifically, when modalities provide complementary information without introducing excessive noise. It further introduces efficiency curves to guide uncertainty-based data acquisition.

The findings show that multimodal models frequently outperform single-modality ones, particularly when the primary modality yields uncertain predictions. Crucially, the analysis reveals that the first 24h of ICU admission represent the critical window for data accumulation, suggesting that earlier collection of specific data types can support more timely and targeted clinical interventions, optimising the use of limited resources.

Chapter 6 (Ablation on subgroup differences in multimodal data fusion) examines subgroup performance differences in multimodal clinical prediction, focusing on cases where demographic information is present both explicitly in tabular inputs and implicitly in imaging data. It uses multi-task learning and the SPLIT framework to test for demographic redundancy (Glocker et al., 2021) and investigates group-dependent miscalibration. The chapter evaluates post-training fairness interventions (Delaney et al., 2024), including per-group decision thresholds with fairness constraints, and analyses under which conditions they are effective and whether these approaches generalise to held-out test data.

The study reveals that explicitly adding demographic features can result in misleading effect sizes due to double counting. It further demonstrates that per-group threshold optimisation, a common fairness intervention, only yields significant advantages under specific conditions of low class separation and high prevalence differences. Additionally, it finds that optimising F1-scores with minimum-recall constraints offers greater stability than accuracy-based constraints, providing a practical framework for ensuring algorithmic fairness in deployment.

Table 1: **Thesis structure.** Summary of chapters that form this thesis.

Chapter	Venue	Research aim	Main contributions
Chapter 2 - Review of multimodal machine learning approaches in healthcare	Published in <i>Information Fusion</i> (2025).	Understand the gaps and open questions in current multimodal research for clinical decision making.	Structured review of multimodal machine learning in healthcare that categorises data modalities, fusion strategies and training paradigms and synthesises evidence from over 50 studies. It shows that performance gains from multimodality are strongly task and data dependent, supported by a consolidated overview of 17 key clinical datasets.
Chapter 3 - Digitising ECG waveforms for use in multimodal pipelines	Published in <i>Computing in Cardiology</i> (2024) and presented at CinC in Karlsruhe, winning the PhysioNet Challenge.	Understand how waveform signals can be reliably reconstructed from paper printouts using a generalisable and robust digitisation model across diverse layouts and image qualities.	Improved ECG digitisation method that outperformed the benchmark by more than twice, enabling the use of imaging-based pipelines, focusing on generalisability.
Chapter 4 - A generalisable approach for multimodal waveform integration	Published in <i>Computing in Cardiology</i> (2023) and presented at CinC in Atlanta, winning the best preprint award.	Understand if spectrogram-based image models can offer a generalisable alternative to expert-designed feature extraction for waveform-based tasks.	New, generalisable, spectrogram-based modelling pipeline as a practical alternative to expert-derived features for waveform analysis.
Chapter 5 - Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders	Published in <i>Information Fusion</i> (2026).	Understand three key questions: Does any fusion method consistently achieve superior performance in multimodal prediction of cardiovascular conditions in the ICU? What is the contribution of each data modality to predictive performance for cardiovascular conditions in the ICU? At what point during a patient's stay in the ICU has sufficient data accumulated to make reliable predictions?	Structured evaluation of multimodal fusion strategies for ICU cardiovascular prediction, systematically comparing early, intermediate and late fusion across imaging, ECG, text and tabular data in MIMIC-IV. It demonstrates that late fusion is most robust under realistic data availability, with largest gains in high-uncertainty cases and within the first 24 hours of ICU admission.
Chapter 6 - Ablation on subgroup differences in multimodal data fusion	To be submitted.	To determine whether multimodal late fusion introduces demographic redundancy and contributes to subgroup performance differences and to evaluate whether post hoc calibration and fairness constraints reduce subgroup disparities and identify when such interventions are expected to help.	Structured analysis of subgroup differences in multimodal diagnostics, showing that per-group threshold calibration only helps under low class separation and large prevalence gaps and that fairness constraints selected on validation data often fail to transfer, with F1-optimised minimum-recall constraints exhibiting greater stability than accuracy-based criteria.

Finally, **Chapter 7 (Discussion)** synthesises the key findings and outlines future research directions for robust, multimodal clinical AI. In Appendix B I reflect on my contributions to PhysioNet.

1.4 Publications

During my DPhil I contributed to 12 publications, including six first-authored papers.

The following papers form the core chapters of this thesis:

- **Krones F**, Marikkar U, Parsons G, Szmul A, Mahdi A. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690 (2025).
- **Krones F**, Walker B, Lyons T, Mahdi A. Combining hough transform and deep learning approaches to reconstruct ECG signals from printouts. *2024 Computing*

in Cardiology (CinC), Karlsruhe, Baden-Württemberg, Germany, 2024, pp. 1-4,

- **Krones F**, Walker B, Parsons G, Lyons T, Mahdi A. Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest. *2023 Computing in Cardiology (CinC)*, Atlanta, GA, USA, 2023, pp. 1-4.
- **Krones F**, Papastylianou T, Qian G, Parsons G, Shukla B, Papież BW, Mahdi A. Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders. *Information Fusion*, 136:104573 (2026).

Beyond the core narrative of the thesis, during my DPhil I contributed to a broader range of studies that explore complementary aspects of medical AI. These publications investigate specific deployment challenges, such as the application of large language models in healthcare, the development of open-access benchmark datasets and the creation of lightweight architectures for low-resource environments:

- **Krones F**, Walker B. From theoretical models to practical deployment: A perspective and case study of opportunities and challenges in AI-driven cardiac auscultation research for low-income settings. *PLOS Digital Health* 2024 4;3(12):e0000437.
- **Krones F**. From pre-training to downstream performance: Does domain-specific pre-training make sense? arXiv:2605.08819.
- Duvieusart B, **Krones F**, Parsons G, Tarassenko L, Papież BW, Mahdi A. Multimodal cardiomegaly classification with image-derived digital biomarkers. In *Annual Conference on Medical Image Understanding and Analysis 2022* (pp. 13-27). Cham: Springer International Publishing.
- Duvieusart B, **Krones F**, Parsons G, Tarassenko L, Papież BW, Mahdi A. Heart and lung segmentations for MIMIC-CXR/MIMIC-CXR-JPG and Montgomery County TB databases (version 1.0.0). *PhysioNet*. RRID:SCR_007345.
- Duvieusart B, **Krones F**, Parsons G, Tarassenko L, Papież BW, Mahdi A. Image-derived cardiomegaly biomarker values for 96K chest X-rays in MIMIC-CXR/MIMIC-CXR-JPG (version 1.0.0). *PhysioNet*. RRID:SCR_007345.

- Walker B, **Krones F**, Kiskin I, Parsons G, Lyons T, Mahdi A. Dual Bayesian ResNet: A deep learning approach to heart murmur detection. In *2022 Computing in Cardiology (CinC)* Sep 4 (Vol. 498, pp. 1-4). IEEE.
- Korgul K, Bean AM, **Krones F**, McCraith R, Mahdi A. Exploring the landscape of large language models in medical question answering: Observations and open questions. arXiv:2310.07225. Presented at *NeurIPS 2024* AIM-FM workshop.
- Amelung D, Bean AM, **Krones F**, Herpertz SC, Parsons G, Mahdi A, Schneider I. How sensitive do we want AI to be? Socio-communicative competencies of large language models in healthcare. Note: Submitted in June 2025.

2 Review of multimodal machine learning approaches in healthcare

This chapter is based on the paper (Krones et al., 2025):

Krones F, Marikkar U, Parsons G, Szmul A, Mahdi A. *Review of multimodal machine learning approaches in healthcare*. **Information Fusion** (2025) 114:102690.

Abstract. Machine learning methods in healthcare have traditionally focused on using data from a single modality, limiting their ability to effectively replicate the clinical practice of integrating multiple sources of information for improved decision making. Clinicians typically rely on a variety of data sources including patients’ demographic information, laboratory data, vital signs and various imaging data modalities to make informed decisions and contextualise their findings. Recent advances in machine learning have facilitated the more efficient incorporation of multimodal data, resulting in applications that better represent the clinician’s approach. Here, we provide an overview of multimodal machine learning approaches in healthcare, encompassing various data modalities commonly used in clinical diagnoses, such as imaging, text, time series and tabular data. We discuss key stages of model development, including pre-training, fine-tuning and evaluation. Additionally, we explore common data fusion approaches used in modelling, highlighting their advantages and performance challenges. An overview is provided of 17 multimodal clinical datasets with detailed description of the specific data modalities used in each dataset. Over 50 studies have been reviewed, with a predominant focus on the integration of imaging and tabular data. While multimodal techniques have shown potential in improving predictive accuracy across many healthcare areas, our review highlights that the effectiveness of a method is contingent upon the specific data and task at hand.

Contents

2.1	Introduction	18
2.2	Data modalities	21
2.3	Model development	31
2.4	Fusion approaches	38
2.5	Multimodal datasets and studies	45
2.6	Conclusions and future research	47

2.1 Introduction

Artificial Intelligence (AI), and machine learning in particular, has radically altered our interactions with the world, fostering rapid advancements in various domains. However, the adoption of machine learning approaches in healthcare has been slower than in other fields despite the increasing pressure on healthcare systems and the urgent demand for high quality, personalised care ([Kirch and Petelle, 2017](#); [Topol, 2019](#)).

Machine learning methods can already outperform humans in some areas, especially when time restrictions apply ([Bejnordi et al., 2017](#); [Topol, 2019](#)) and their integration into clinical workflows presents a significant opportunity to improve healthcare and alleviate challenging resource constraints ([Bartoletti, 2019](#)).

Advanced machine learning techniques have supported the integration of multimodal healthcare data, allowing the creation of methods that mimic clinical practice by incorporating diverse information sources into decision-making processes. Clinicians often consider multiple data sources when assessing a patient’s condition and determining appropriate treatment, taking advantage of the additional information and context this approach supports. For example, a study of radiologists found that 85% deemed the clinical context to be crucial for the interpretation of radiological examinations ([Leslie et al., 2000](#)). Patients’ demographic data, medical history, laboratory data and vital signs can all be used by physicians to contextualise their assessments of medical images within the broader clinical picture. In response to this, an increasing volume of AI research in healthcare has focused on the use of multimodal data, aiming to better emulate clinicians’ approaches and improve overall performance ([Soenksen et al., 2022](#)).

The field of machine learning in healthcare and medicine has been extensively explored through various survey studies, covering a range of topics. These include general applications of deep learning in medicine ([Akay and Hess, 2019](#); [Esteva et al., 2019](#); [Fatima et al., 2017](#); [Piccialli et al., 2021](#); [Shamshirband et al., 2021](#); [Zhang et al., 2018](#)) and image-focused methodologies ([Litjens et al., 2017](#); [Shen et al., 2017](#); [Wang et al.,](#)

2021a). Area specific reviews, for example, include those focused on the chest region (Bizopoulos and Koutsouris, 2018; Çalli et al., 2021; Krittanawong et al., 2019). In the context of multimodal applications, some papers have reviewed the field in general (Acosta et al., 2022; Cui et al., 2022; Kline et al., 2022; Pei et al., 2023; Salvi et al., 2023), while others have concentrated on particular medical fields such as oncology (Lipkova et al., 2022) and cardiology (Amal et al., 2022). Additional investigations have focused on the fusion of the different medical imaging modalities, including MRI, CT and PET (Azam et al., 2022; Basu et al., 2024; Sebastian and King, 2021; Zhou et al., 2023b). Furthermore, some reviews have compared models, architectures and optimisation strategies, either in general (Gao et al., 2020; Ramachandram and Taylor, 2017) or within specific healthcare contexts (Ayesha et al., 2021; Behrad and Abadeh, 2022; Heiliger et al., 2022; Huang et al., 2020b,c; Stahlschmidt et al., 2022; Xu et al., 2021). Other reviews have focused on the emerging field of self-supervised learning, following its development and application (Fei et al., 2022; Krishnan et al., 2022; Shurrab and Duwairi, 2022; Weng et al., 2024; Wójcik, 2022).

While some reviews on multimodal machine learning approaches in healthcare do exist, our aim here is two-fold. We provide an elementary and self-contained introduction to data fusion techniques in healthcare, carefully explaining the fundamental concepts for each technique. We subsequently review the latest developments in this rapidly developing field in particular, focusing on available multimodal datasets and clinical applications. Unlike many previous reviews that primarily focus on specific medical conditions and the performance of multimodal methods within those contexts, this review addresses a wider spectrum of multimodal machine learning applications in healthcare. It covers diverse data modalities, fusion approaches and stages of model development. Furthermore, we offer a comprehensive overview of multimodal clinical datasets and studies, highlighting the significance of aligning multimodal machine learning techniques with clinical practices—an aspect less emphasised in earlier reviews. Specifically, this review covers commonly used data types in clinical applications, notably imaging, text, time-series and tabular data, as well as techniques for fusing these different

data modalities. Examples of how multimodal data can be incorporated in practical applications are given through an overview of a number of clinical datasets that were included in over 50 recent studies. Our focus was on studies including images as they offer rich visual information that is crucial for many multimodal integration tasks, making them a central element in current research.

Overall, this study addressed the following questions:

RQ 1: *Is there a common approach to compare fusion strategies?*

RQ 2: *What are the most common ways to combine multiple data sources in healthcare applications?*

RQ 3: *Which data sources are available for clinical multimodal research?*

The paper is organised as follows. Section 2.2 provides an overview of four key data modalities commonly used in clinical practice. Section 2.3 describes the main development stages of deep learning models including pre-training and fine-tuning. Section 2.4 describes different fusion approaches, at both modality- and feature-level. Section 2.5 gives a comprehensive overview of recent studies using machine learning data fusion techniques in clinical contexts as well as a number of multimodal datasets used to train such models. Finally, Section 2.6 examines the opportunities and challenges associated with multimodal approaches in healthcare and outlines future research needs.

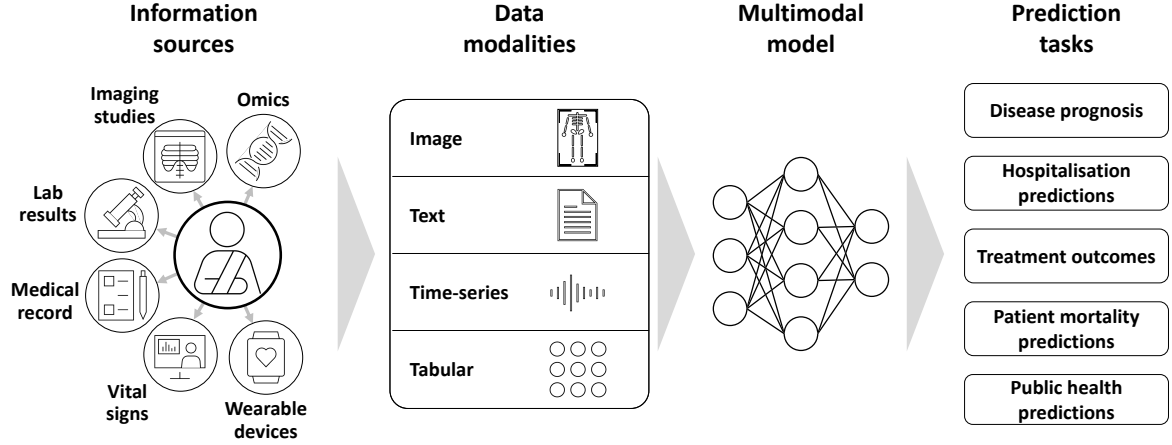


Figure 1: **Clinical data modalities and prediction tasks.** Distinct data modalities play pivotal roles in clinical decision-making: imaging data, text data, time-series data and tabular data. All are used for various clinical predictions tasks. Medical prediction tasks in clinical practice involve leveraging machine learning models and algorithms to forecast future clinical outcomes based on existing patient data. They play a crucial role in the decision-making process for diagnosis, prognosis and treatment.

The main contributions of this chapter are:

- **Gap identification of systematic approaches for fusion model development and evaluation:** The paper highlights the lack of standardised methodologies for building, comparing and evaluating fusion models in healthcare, emphasising the need for more structured frameworks.
- **Comprehensive review of multimodal fusion strategies and future trends:** The paper reviews modality-level fusion techniques for healthcare applications and discusses emerging trends, along with deployment challenges.
- **Analysis of multimodal datasets for healthcare applications:** The study examines the availability of clinical multimodal datasets, highlighting the shortage of large, publicly available multimodal datasets.

2.2 Data modalities

Multiple data modalities find application in clinical practice (Duvieusart et al., 2022; Qayyum et al., 2020) as illustrated in Figure 1 and described in Table 2. These include imaging data (e.g. X-rays), text data (e.g. radiology reports), time-series data (e.g. vital signs) and cross-sectional or panel tabular data (e.g. metadata). Recognising the critical role of data selection in multimodal machine learning, we provide an overview of the key data modalities used in modelling studies.

Table 2: **Clinical data modalities.** The table summarises four common clinical data types, giving representative examples, typical storage formats, key data properties, practical challenges (e.g. storage requirements, missingness, annotation effort, integration across systems) and suitable analysis approaches. The listed methods are illustrative rather than exhaustive and highlight how differences in structure and sampling drive modality-specific pre-processing and model choice in multimodal clinical pipelines.

Data type	Examples	Common formats	Key characteristics	Challenges	Suitable analysis methods
Image	X-Ray, CT, MRI, ultrasound, dermoscopic images	DICOM, PNG, JPEG, Nifti	High-dimensional data providing detailed visual information on anatomical and pathological states; contains spatial information.	Requires large storage space; interpretation demands significant expertise; high computational cost for processing; ensuring privacy and anonymisation.	CNNs and transformers for image classification, segmentation and improvement; transfer learning for leveraging pre-trained models on large datasets.
Text	Clinical notes, EHRs, radiology reports	Plain text (.txt), PDF, XML, JSON	Unstructured or semi-structured; rich in clinical information; includes medical jargon, abbreviations and patient-specific nuances.	Requires NLP to extract meaningful information; variability in language and structure; privacy concerns.	NLP techniques to extract useful information, such as deep learning models including RNNs and transformers for text classification and summarisation.
Time-series	ECG readings, continuous vital signs monitoring (heart rate, blood pressure), wearable device data	CSV, relational databases (SQL)	Sequential data, recorded over a period; can be regular or irregular intervals; often requires real-time analysis.	Handling of large volumes of data; dealing with missing values and noise; time synchronisation across different sources.	Time-series analysis techniques (e.g. ARIMA, seasonal decomposition), LSTM networks for prediction and anomaly detection; Fourier transforms for signal processing.
Tabular	Patient demographics, laboratory test results, medication records	CSV, Excel, relational databases (SQL), Parquet	Structured data with rows and columns; can be numeric, categorical, or a mix; amenable to traditional statistical analysis.	Missing values, imbalanced datasets, high dimensionality; integration with other data types for comprehensive analysis.	Statistical models (e.g. logistic regression for classification tasks), machine learning algorithms (e.g. random forests, gradient boosting machines), deep learning for complex patterns, feature selection and dimensionality reduction techniques (e.g. PCA).

2.2.1 Imaging data

Most medical imaging data are stored as 2D image slices in the Digital Imaging and Communications in Medicine (DICOM) format after acquisition (Bidgood Jr et al., 1997). This includes patients’ metadata, details regarding the imaging procedure, information about the device used for image acquisition and imaging protocol settings. Medical 3D volumetric images are typically constructed from a stack of 2D slices with a specified thickness, representing a specific region of interest within the body. These individual slices can be processed and analysed separately (as 2D images) or collectively (as 3D volumes) to extract valuable information.

When preparing medical imaging data for machine learning, data in the DICOM format are either converted to commonly used image formats, such as JPEG or PNG (Tan, 2006), or to the Neuroimaging Informatics Technology Initiative (NifTi) format (Poldrack et al., 2019), a dedicated medical image analysis format where, along with the image, only essential metadata are stored in the header.

X-ray imaging. Medical X-ray imaging is a widely available and inexpensive 2D imaging technique. Almost half of the 43.3 million imaging scans in England in 2022 were X-ray images, making it the most commonly used medical imaging modality (NHS, 2016). X-ray imaging operates on the principle of differential attenuation of X-rays through various types of body tissues (Dendy and Heaton, 2011). X-rays are inherently 2D grey-scale radiograph images. Conventional radiographs identify five levels of attenuation: air, fat, soft tissue, bone and metal. In radiographs, air appears the darkest because its low density allows most X-rays to pass through, whereas metal, being significantly denser, appears bright white due to its absorption of the majority of the energy of the X-ray beam. Fat tissue, soft tissue and bone manifest as varying shades of grey, with fat being darker than soft tissues and bone lighter (Brant and Helms, 2012).

In machine learning applications, images are resized to specific fixed dimensions, such as 224×224 (e.g. the original VGGNet was trained with a resolution of 224 pixels (Simonyan and Zisserman, 2014)), 256×256 or 512×512 pixels. Subsequently, depending on the model, grey-scale X-ray images are either maintained as single-channel images or transformed into three-channel images by replicating the existing channel (Duvieusart et al., 2022).

While bone imaging remains the most common application for X-ray scans, their utility extends to various other areas of the body, encompassing radiography, fluoroscopy and mammography (primarily used for breast cancer screening) (Göttsche and Jørgensen, 2013; Martensen, 2013). Since X-rays lack sufficient spatial depth information for unequivocal diagnosis, they are frequently used as a screening tool. Many recent

studies have used X-rays in cardiology prediction tasks, such as cardiomegaly diagnosis (Duvieusart et al., 2022; Palepu and Beam, 2022), pulmonary oedema assessment (Chauhan et al., 2020) or accessory conduction pathway analysis (Nishimori et al., 2021).

Computed Tomography (CT). Computed Tomography (CT) scans provide detailed cross-sectional images of a human body (Buzug, 2011). These scans generate 3D image volumes by reconstructing multiple subsequent 2D slices from radiographic projections acquired from multiple angles. CT scans are depicted in Hounsfield Units (HU), which directly correspond to signal attenuation caused by tissue density with respect to water (Goldman, 2007).

As a popular choice for medical diagnostics, CT scans offer high-resolution imaging, wide availability, cost-effectiveness and speed. However, they expose patients to ionising radiation and have limitations in distinguishing soft tissues (Goldman, 2007).

CT is a versatile imaging technique primarily used to identify structural abnormalities, detect tumors, diagnose heart diseases and image the head for various neurological conditions (Adam et al., 2020; Bremner, 2005; Schoepf et al., 2007). It is often used in respiratory (Doğan et al., 2015) and cardiological studies (Fang et al., 2021; Samak et al., 2020; Xu et al., 2020), in cancer diagnosis (Wiener et al., 2013) and treatment planning (Battista et al., 1980; Brunelli et al., 2009).

Magnetic Resonance Imaging (MRI). Unlike the previously discussed imaging techniques, Magnetic Resonance Imaging (MRI) is a non-ionising method (Hashemi et al., 2010). The patient is placed in a strong magnetic field, which aligns the protons’ magnetic moments (spin) of their body with the field. Short radio-frequency pulses disturb the alignment of these protons and then realign with the magnetic field. MRI measures magnetisation in both longitudinal and transverse directions, enabling tissue-specific reconstructions (Grover et al., 2015).

MRI maintains a high signal-to-noise ratio and provides a detailed view in two directions (Stadler et al., 2007). However, it may lead to aliasing artifacts in the orthogonal direction, especially in 2D magnetic resonance acquisition protocols. To address such

artifacts, magnetic resonance data may require anti-aliasing pre-processing steps, such as filtering or the application of dedicated machine learning models (Stadler et al., 2007).

This technique works well for providing detailed visuals of soft tissues organs and internal structures without employing ionising radiation (Frisoni et al., 2010; Guermazi et al., 2013). It is extensively used to study brain disorders, including Alzheimer’s disease (Parisot et al., 2018; Pölsterl et al., 2021), multiple sclerosis (Yoo et al., 2019) and Parkinson’s disease (Ryman and Poston, 2020).

Nuclear medicine imaging techniques. Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) are nuclear medicine imaging techniques that detect gamma photons emitted from radioactive tracers in the body, offering insights into metabolic activity (Israel et al., 2019) and blood flow/function (Mullani et al., 2008).

In practice, these techniques generate dozens to hundreds of 2D slices acquired per scan (Dwivedi et al., 2022). The slices are evenly spaced, ensuring consistent distance within the scan. The intensities within the images have a relative value and each scan is often accompanied by a paired attenuation correction CT image. The main limitations of these imaging modalities include long acquisition times compared to CT and usually lower resolution than CT or MRI.

Nuclear medicine imaging techniques are important for diagnosing diseases, monitoring treatments and studying health at the molecular level (Braman et al., 2021; Dwivedi et al., 2022). PET can be useful in cancer diagnosis and for neurological conditions (Duclos et al., 2021) while SPECT can improve staging, prognosis determination and treatment planning (Israel et al., 2019). These techniques are often paired with CT and MRI, providing complementary insights in tasks such as Alzheimer’s prediction with MRI (El-Sappagh et al., 2020; Suk et al., 2014), lung cancer prediction with CT (Hyun et al., 2019) or tumor segmentation tasks using both modalities (Guo et al., 2019).

Ultrasound. Ultrasound imaging uses acoustic energy with frequencies above 20 kHz to visualise body structures (Woo, 2002). The data typically consist of a series of

2D frames (Carovac et al., 2011); however recent advancements now enable 3D and 4D (real-time 3D) imaging (Merz and Abramowicz, 2012). Image resolution varies depends on transducer frequency; higher frequencies provide greater resolution but less depth penetration and so are more useful for superficial structures while lower frequencies provide greater depth at the cost of resolution. Image intensities correspond to tissue echogenicity (i.e. the ability to reflect sound waves), with different shades of grey representing different tissue densities (Woo, 2002). Doppler techniques are often incorporated to visualise blood flow and assess velocity, providing helpful colour overlays on the grey-scale images (Carovac et al., 2011).

Prior to applying machine learning techniques, ultrasound data are pre-selected, as the spatial position or orientation is not immediately apparent and relies on the operator’s approach (Brattain et al., 2018). This often involves frame selection and identifying the region of interest (ROI) which can be achieved through either manual or automated methods. Ultrasound data inherently exhibit speckling, a characteristic that may require compensation during pre-processing (Karaoglu et al., 2022).

The non-invasiveness of ultrasound and lack of ionising radiation make it a favoured choice for obstetrics and gynecology (Merz and Abramowicz, 2012). It can be used to evaluate heart function, monitor blood flow (Brattain et al., 2018) and review many organs (e.g. liver, kidneys) for potential issues (Carovac et al., 2011). Beyond these applications, ultrasound plays an important role in tracking disease progression and guiding precise surgical procedures (Carovac et al., 2011).

Dermoscopic images. Dermoscopy is a non-invasive imaging technique that captures high-resolution images of the skin surface and is primarily used for the early detection of skin cancers such as melanoma (Vestergaard et al., 2008). By employing polarised or non-polarised light, dermoscopy can visualise skin structures that are otherwise invisible to the naked eye.

Dermoscopic images can be analysed using computer-aided diagnostic tools which leverage machine learning approaches to assist dermatologists in making more accurate

and efficient diagnoses. Before presenting images to a machine learning model, regions of interest are usually segmented to remove redundant information (Kawahara et al., 2018). The images also need to be cleaned of unwanted artefacts, such as hair, gel bubbles and ink marks (Iqbal et al., 2021). As they are photographic images, they usually have three channels (RGB) like natural images, compared to most medical imaging which has only one channel for intensity.

Dermoscopic images are primarily employed in dermatology to visualise subsurface skin structures in the epidermis and dermis, aiding in the diagnosis and monitoring of skin lesions and tumors (Gessert et al., 2020; Yap et al., 2018). The technique has proven valuable for distinguishing between benign and malignant lesions, improving diagnostic accuracy and reducing the need for invasive procedures like biopsies (Kittler et al., 2002).

2.2.2 Text data

The text modality constitutes a fundamental and extensively used form of medical information in healthcare (Bean et al., 2024; Mustafa and Rahimi Azghadi, 2021; Spasic et al., 2020). Various patient-specific text modalities exist, many of which are common in clinical practice. These include procedure notes, comprehensive clinical records generated by healthcare providers, that include progress reports and consultation notes. Prescription notes comprise detailed instructions for medication regimens, specifying drug names, dosages and usage guidelines (Li et al., 2015). Medical discharge notes, critical in patient care transitions, document a patient’s hospitalisation experience, including vital information such as diagnoses, treatments and post-discharge plans (Johnson et al., 2023a). Referral letters serve as communication bridges between healthcare providers, containing pertinent patient details and the rationale behind referrals. Radiology notes offer descriptions of findings extracted from diagnostic imaging studies, encompassing X-rays, CT scans and MRIs (Casey et al., 2021; Duvieusart et al., 2022; Huang et al., 2021). Beyond patient-specific data, the text modality also encompasses general medical knowledge drawn from diverse sources, including medical journals, literature, medical websites and pharmaceutical labels (Xie et al., 2021).

Natural Language Processing (NLP) techniques have been used to extract relevant information from medical records (Locke et al., 2021; Sheikhalishahi et al., 2019). Once this information is extracted it can be converted into a structured format that can then be used in machine learning models. For traditional NLP methods, one of the key challenges is the need for subject matter experts to label features of interest in text sources, a time-consuming task. Various approaches have been used to address this bottleneck, such as active learning techniques to prioritise text for labelling (Chen et al., 2015). Transfer learning, data augmentation techniques and the use of synthetic clinical notes offer further potential in this area (Walonoski et al., 2018).

2.2.3 Time-series data

Time-series data refers to data points that are collected over consistent intervals of time. One common example is the Electrocardiogram (ECG), which traces the electrical activity of the heart over a short period of time, providing insights into heart rhythm and potential abnormalities (Ahsan and Siddique, 2022). Electroencephalography (EEG) records brain activity, enabling clinicians to detect neurological disorders and monitor treatment responses (Zheng et al., 2021). Foetal monitoring allows for the surveillance of a foetus’s well-being during pregnancy and labour, highlighting any distress or deviations through the consistent recording of heart rate and other variables (Freeman et al., 2012). Intracranial pressure monitoring in patients with conditions such as traumatic brain injury can help ensure that brain pressure remains within safe limits (Czosnyka and Pickard, 2004). Other modalities include respiratory monitoring, used in critical care or post-operative contexts (Nicolò et al., 2020); oximetry, which provides real-time data on blood oxygen levels (Luks and Swenson, 2020); and continuous blood pressure monitoring for sustained cardiovascular assessment (Armitage et al., 2023). The continuous nature of these modalities offers clinicians a dynamic view of physiological parameters, enabling early intervention and the evaluation of therapeutic outcomes.

Time-series data typically undergo preprocessing to remove noise and outliers and to address issues such as missing data, sparsity and irregular sampling (Che et al., 2018).

Following this, features are often extracted based on expert knowledge to ensure their relevance and accuracy (Zabihi et al., 2023). Beyond this expert-guided approach, several computational methods, including Fourier transformations, are employed. When coupled with machine learning techniques, these methods prove effective in feature extraction from time-series data (Walker et al., 2022). Time dependencies can be managed during the feature extraction phase or a subsequent step, for instance, by employing LSTM to capture dependencies over time (Zheng et al., 2021).

Time-series data are often used for detecting early signs of patient deterioration or recovery, predicting outcomes and determining appropriate interventions and treatment plans. For example, the presence of rising inflammatory markers and a fever may prompt a change in antibiotic regimen for a patient being treated for infection (Watkins and Lemonovich, 2010), while the resolution of previously abnormal blood tests may prompt a consideration of hospital discharge. Some studies have integrated audio data for various disease prediction tasks (Ceccarelli and Mahmoud, 2022; Salekin et al., 2021; Walker et al., 2022), while others have used vital signs, such as heart rate recordings and blood pressure measurements, in their models (Duvieusart et al., 2022; El-Sappagh et al., 2020; Grant et al., 2021).

2.2.4 Tabular data

Tabular data refers to structured data that can be represented in a table format with rows and columns, generally containing one data point per feature per subject. This differs from time-series data which, although often stored in tabular format, capture continuous information over time for each subject. Tabular data offer a snapshot-like overview while time-series data provide a dynamic view of physiological or clinical variables. In healthcare, the systematic organisation and analysis of various forms of patient information often uses tabular data. For example, demographic data, such as a patient’s age, sex and ethnicity, is often provided in tabular format and can provide essential context for individualised care (Duvieusart et al., 2022; Xu et al., 2020). Health scoring systems, including standardised assessments like disease severity scores (e.g.

APACHE II (Knaus et al., 1985)), pain scales (e.g. KOOS (Pierson et al., 2021)) and quality-of-life indices (e.g. EQ-5D (Herdman et al., 2011)) can offer quantifiable insights into a patient’s condition and treatment efficacy. Clinical laboratory data such as blood chemistry analysis are another very common example and can be used to add granularity to assessments of patient condition and inform prognosis (Mayne et al., 2024; Xu et al., 2020). Pharmaceutical data, such as drug dosages, formulations and pharmacological properties, are often found in tabular format and are vital for medication management and research (Samak et al., 2020). Beyond these core elements, additional layers of tabular data can include insurance and billing information, care plans, social and behavioural data, medical history, medical inventory, clinical outcomes and follow-up data (Johnson et al., 2020). Other data modalities such as time-series data can often be aggregated and presented in tabular form for summary analysis (Duvieusart et al., 2022).

Tabular data provide a foundation for predictive models and decision support systems. These data are predominantly analysed using supervised techniques, such as linear and logistic regression (Pölsterl et al., 2021), decision trees and ensembles of them, like XGBoost and random forests (Krones et al., 2023), as well as support vector machines (Nie et al., 2019). Similar to time-series data, feature engineering emerges as an essential aspect for tabular data. Methods for feature selection such as correlation-based selection and feature extraction techniques such as principal component analysis serve to diminish data dimensionality, enhancing model performance (Ayesha et al., 2021). Deep learning approaches have also made inroads into tabular data analysis, with models such as TabNet (Arik and Pfister, 2021) and 1D-CNNs (Shwartz-Ziv and Armon, 2022).

Tabular data play a significant role in a broad range of medical tasks, including but not limited to, disease prediction (Pölsterl et al., 2021), patient risk evaluation (Bagheri et al., 2020) and outcome forecasting (Vanguri et al., 2022). For instance, tabular data was used with MRI for the detection of prostate cancer (Reda et al., 2018); with demographic and audio data for heart murmur detection (Walker et al., 2022); with demographics, images and time-series data for Alzheimer’s diagnosis (El-Sappagh et al.,

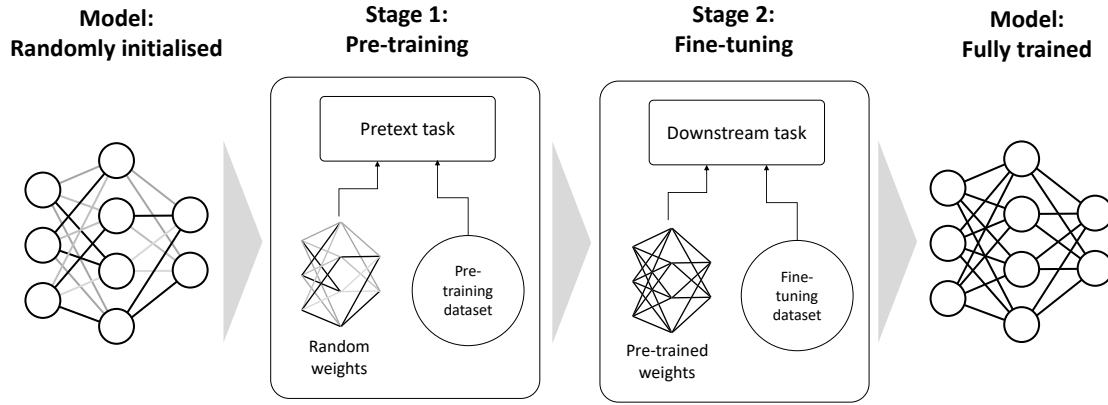


Figure 2: **Model development.** After pre-training the model starting from random weights, the model weights are fine-tuned on the target domain (e.g. medical images) and the model architecture is adjusted to the downstream target task (e.g. classification).

2020); and with genomic data as well as CT and pathological images for predicting lung cancer response to immunotherapy (Vanguri et al., 2022).

2.3 Model development

Developing a multimodal deep learning framework in healthcare follows the same steps as most other machine learning approaches and consists of problem formulation, data pre-processing, model training and model evaluation (IBM, 2021). What makes the deep learning process unique is the notion of pre-training and fine-tuning (Goodfellow et al., 2016), which separates the training procedure into two stages (Figure 2). Pre-training a model aims for a model to understand salient concepts in data, which is usually task-agnostic (Section 2.3.2). Fine-tuning on the other hand transforms these concepts into meaningful information that can be interpreted by humans (Section 2.3.3).

This section will focus on those two training stages—pre-training and fine-tuning—of the deep neural network model, transforming it from an initial random state to a learned state capable of executing specific tasks in downstream applications. In addition to the training stages, we will discuss the data pre-processing step, which is a prerequisite to be carried out whenever training is performed as well as the model evaluation, which is performed after training but before deployment to the real-world.

2.3.1 Data pre-processing

Data pre-processing serves as a critical phase in machine learning pipelines, transforming raw and unstructured data into a format optimised for subsequent model training (Maharana et al., 2022; Nawi et al., 2013). This transformation typically follows data collection and precedes the training of machine learning models.

The process encompasses data cleaning, integration, transformation and reduction. These steps aim to handle missing values, remove inconsistencies and convert categorical variables into numerical form (Little and Rubin, 2019; Maharana et al., 2022). By enhancing data quality, pre-processing facilitates better pattern recognition and relationships within the data, ultimately contributing to improved model performance, reduced overfitting and more reliable and interpretable results.

In *data cleaning*, the aim is to impute missing values and remove duplicates from the raw dataset (Barnard and Meng, 1999). In some cases, outliers are removed ensuring that these outliers do not contain information that represents a meaningful edge case. For example, a fully blacked-out image or a blank text file can be considered to be an outlier as they do not contain any useful information (Yu et al., 2002).

Data integration involves combining and merging data from multiple sources or datasets to create a unified and comprehensive dataset (Duvieusart et al., 2022). This process aims to facilitate more comprehensive analysis, modeling and decision-making. For example, integrating multimodal data from multiple body parts to train a proposed framework can be especially useful for a model to be more robust towards out-of-distribution data (i.e. data that are fundamentally different from the data on which a machine learning model has been trained) (Chen et al., 2022).

Data transformation involves altering the format, structure or values of data to make them more suitable for analysis or modeling. This may involve obtaining vector embeddings for raw text data (Devlin et al., 2018) or resizing images for image data (Krizhevsky et al., 2012). This process aims to improve data quality and prepare the data for specific machine learning algorithms. It often involves data scaling (bringing numerical

variables to a similar scale or range) and data normalisation to allow for more efficient training (Ioffe and Szegedy, 2015). When considering inputs from multiple modalities, normalisation enables the model to treat each data stream fairly, preventing any single modality from exerting undue influence on the overall input.

Finally, *data reduction*, which consists of feature selection or dimensionality reduction, is carried out in cases where the data remain too noisy after cleaning and normalisation or where the dimensionality of the data points may be too high for the model. Dimensionality reduction can be performed using techniques like principle component analysis (Pearson, 1901). Data reduction steps convert the data into a trainable state and thereby enable neural networks to interpret the data in an optimal manner.

2.3.2 Stage 1: Model pre-training

Model pre-training (or *representation learning*) aims to build strong modality-specific or joint representations of the input data by training a neural network to perform a specific *pretext task* (i.e. a task involving a loss function that allows the network to learn these representations) (Shurrab and Duwairi, 2022). The representations of the input data are obtained by training the model on large datasets, such as ImageNet (Deng et al., 2009) for image-based tasks, BookCorpus (Devlin et al., 2018) for text-based tasks or MIMIC (Johnson et al., 2023a) for medical images. These provide a strong initialisation point for fine-tuning networks, typically on a much smaller dataset.

Pre-training can itself contain multiple stages. For example, a model that has already been trained on natural images can be further pre-trained on a medical domain-specific dataset to provide an even stronger initialisation point for later fine-tuning on another dataset in the medical domain (Kalapos and Gyires-Tóth, 2022). This intermediate step is sometimes interpreted as a fine-tuning step. However, in this review we define fine-tuning as updating the model only for the required downstream task and pre-training as the stage that provides the initialisation point for fine-tuning the model.

The pre-training learning strategy can also be further categorised into supervised and

unsupervised (or self-supervised) learning.

Supervised pre-training. In *supervised learning* a model is trained on a large dataset that contains labels and uses the associated label as a supervisory signal (Hastie et al., 2009). The pretext task in supervised learning is a simple classification or regression task based on the label. A common example is the training of deep networks on ImageNet-1k (Deng et al., 2009) using the given class labels per data-point to perform a classification task. As the dataset is large, the model will be able to generalise a given class based on the multiple examples the model sees under that class. In a multimodal setting, a signal from a single modality may act as a supervisory signal for another modality from the same data point (Chen et al., 2023; Radford et al., 2021).

Supervised pre-training has proven to be strong on very large datasets such as ImageNet-21k (Ridnik et al., 2021), mainly due to the generalisation achieved through the exposure to many examples. However, models trained using self-supervised learning have shown to be able to outperform models employing supervised learning when trained on smaller datasets such as ImageNet-1k (Atito et al., 2021). This is because the learning capacity of supervised learning is limited, as it is constrained by the information contained in the attached supervisory signals (labels). Moreover, supervised pre-training is also disadvantageous due to the labour costs incurred through the need to label large scale datasets (Krishnan et al., 2022).

Self-supervised pre-training. To mitigate the disadvantages of supervised pre-training, *self-supervised learning* aims to train neural networks to learn data representations by leveraging the inherent properties of the data (Ericsson et al., 2022). Hence, the data do not need to be associated with a specific label; consequently, the learned representations can be more generalisable (Azizi et al., 2021; Ericsson et al., 2021; Hendrycks et al., 2019b).

There exists a large number of pretext tasks for self-supervised learning that do not involve pre-defined labels, with commonalities and differences (Krishnan et al., 2022). Here we have grouped them into two categories: discriminative and restorative. In

discriminative tasks, the loss function of the pretext task guides the model to cluster similar data points of the dataset together (Haghighi et al., 2022). One way this can be performed is by contrastive learning (Sowrirajan et al., 2021), where specific loss functions enable association between similar observations and disassociation between dissimilar ones. *Restorative tasks* aim to learn representations of data by forcing the neural network to reconstruct a given raw data point based on some reduced or corrupted representation of that data point (Haghighi et al., 2022). For example the language model GPT-3 (Brown et al., 2020) was trained to predict the next word in partially masked sentences, while image-based model have been trained to reconstruct an image when provided with a distorted one (Atito et al., 2022).

2.3.3 Stage 2: Model fine-tuning

Once pre-training has been performed, the trained model is expected to hold rich information about the semantic concepts in the pre-training dataset (Hendrycks et al., 2019a). This makes it possible to use the learned parameters (weights) from that model for multiple downstream tasks. The literature has shown that if the distribution of the downstream dataset aligns closely with the pre-training dataset, a model pre-trained on medical images is likely to outperform models pre-trained on natural images during fine-tuning, provided the datasets are of comparable size (Ma et al., 2022). Unlike pre-training, *fine-tuning* does not involve pretext tasks and instead is carried out for a specific downstream task that is formulated based on the problem at hand. This usually involves two things: adjusting the model to the target task structure (e.g. for a classification problem, adjusting the last layer of the model); and training the model on the target data, which can either mean training all weights or only weights of specific layers. For multimodal data, modality-specific pre-trained models can be combined and then fine-tuned for a downstream task that involves multimodal data inputs. Together, pre-training a model on a source task and then fine-tuning it on a related target task is referred to as *transfer learning*, a strategy that is especially helpful in healthcare, where labelled data are often scarce (Walker et al., 2022). We describe its main forms,

homogeneous and heterogeneous transfer learning, in more detail in [Section A.1.1](#).

2.3.4 Model evaluation

Model evaluation is carried out prior to model deployment to real-world scenarios. It aims to test the model under various circumstances to ensure its effectiveness, reliability and generalisability ([Beede et al., 2020](#); [Widner et al., 2023](#)).

Deployment may expose the model to a substantial amount of out-of-distribution data that does not align with the distribution of the training data and so evaluating the performance of the model across various patient subgroups, diverse conditions and different geographical regions is essential to test its reliability and generalisability ([Azizi et al., 2022](#)). This becomes particularly critical in a multi-site setup where data can significantly vary across healthcare facilities and populations ([Beede et al., 2020](#); [Tran et al., 2022](#); [Widner et al., 2023](#)). Randomised controlled trials would form an ideal testing scenario, however, these have only been conducted by a handful of studies and are challenging to conduct ([Han et al., 2023](#)). In machine learning, the concerns mentioned above can be categorised as the following key measures: robustness to data noise, interpretability and generalisability.

Robustness to data noise refers to the ability of a machine learning model to maintain stable and accurate performance even when the input data contains irrelevant or erroneous information, often referred to as ‘noise’ ([Tran et al., 2022](#)). It is related to an understanding of how much noise affects predictions (i.e. the uncertainty ([Abdar et al., 2021](#))), which subsequently contributes to clinical decision-making ([Chua et al., 2023](#)). There are various approaches to evaluating uncertainty for deep neural networks including Bayesian and ensemble methods and test-time augmentations of data ([Gawlikowski et al., 2023](#)).

Interpretability and explainability refers to the degree to which the predictions and decisions of a machine learning model can be understood and explained by humans ([Reyes et al., 2020](#)). It involves making complex and often opaque machine learning

models more transparent and comprehensible to users, domain experts and stakeholders. Interpretable models provide insights into how they arrive at specific outcomes, which variables or features are influential and why certain decisions are made. Interpretability is crucial in healthcare as it helps build trust and ensures that the behaviour of the model aligns with ethical and legal standards (Voigt and Von dem Bussche, 2017). The broader field that develops such methods is Explainable Artificial Intelligence (XAI), which provides mostly post-hoc techniques to make model decisions more transparent and to show which inputs drive a prediction (Tjoa and Guan, 2020; Watson, 2021). Beyond saliency maps and attention, common examples include gradient-based attribution methods such as Grad-CAM (Selvaraju et al., 2017). Such post-hoc explanations are, however, not always faithful to the underlying model and can be unstable or misleading, so the explanations themselves should be tested rather than trusted by default (Watson, 2021). In multimodal settings, explainability is harder, as methods have to be aligned across heterogeneous data types, a point we return to in Section 2.6.

Generalisability of a machine learning model refers to the ability of a trained model to perform effectively and accurately on new, unseen data or with patient populations that were not part of the training dataset of the model. In the context of pre-training models, generalisability can be evaluated by performing zero-shot-predictions, where predictions are made on new datasets that the model has not seen before (Tran et al., 2022). Achieving strong generalisability is vital in healthcare applications if models are to provide valuable support across clinical environments and patient demographics while maintaining high performance (Widner et al., 2023). Common methods for assessing the generalisability of machine learning models include evaluating on a test set of unseen data (Azizi et al., 2022), testing on diverse datasets from various institutions (Azizi et al., 2022), conducting clinical validation and expert review (Widner et al., 2023).

2.4 Fusion approaches

Here we describe basic concepts behind data fusion and multimodal deep learning. Comparing fusion techniques in multimodal machine learning is not straightforward, being dependent on the data type and the specific task at hand. However, important characteristics to consider include accuracy and predictive performance, generalisability across different datasets and conditions, computational efficiency and interpretability of the output of the model. As the performance is data and context dependent, we opted to limit ourselves to a qualitative comparison in this study, but investigated different fusion approaches in subsequent research for specific use cases (Krones et al., 2023; Walker et al., 2022). More general quantitative comparison attempts can be found in recent literature, e.g. for general image fusion (such as of natural images) (Zhang et al., 2021), for medical image fusion (such as of CT and MRI scans) (Hermessi et al., 2021; Tirupal et al., 2021), for neural architecture search for electronic health records in favour of hybrid approaches (Xu et al., 2021) or for the fusion of CT images with electronic health records in favour of late fusion (Huang et al., 2020c).

2.4.1 Modality-level fusion

Numerous taxonomies exist in the scientific literature for describing fusion approaches. Depending on how modalities are combined for a final prediction, the following techniques can be distinguished (Figure 3): early fusion, intermediate fusion, late fusion and mixed approaches (Qiu et al., 2022).

Early fusion. In *early fusion* (Huang et al., 2020b,c; Kline et al., 2022) (also referred to as *data-level fusion* (Ramachandram and Taylor, 2017) or *feature-level fusion* (Ayesha et al., 2021; Dolly and Nisa, 2019; Hermessi et al., 2021)) multiple input modalities are combined before training a single machine learning model (see Figure 3(a)). The data can be used in its raw form or may require various steps of feature extraction. This can range from simple aggregation methods to the use of separate models. The method of combining features also depends on the specific modalities and models involved. For

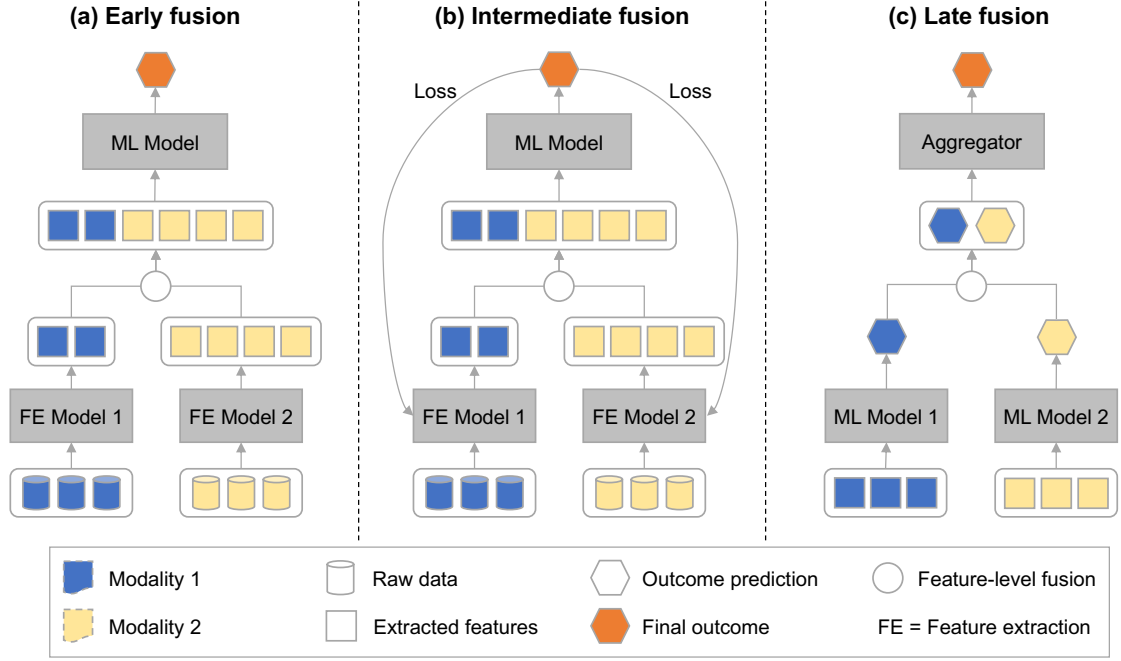


Figure 3: **Data fusion architectures:** (a) Early fusion combines raw features or extracted features before passing them into the final model. The feature extraction is optional; (b) Intermediate fusion concatenates features extracted from the original data using an integrated modelling approach where the loss is back-propagated through the whole model; (c) In late fusion the predictions or features are generated by multiple models and aggregated after their individual processing.

instance, time-series data may need to be aggregated before use with models such as XGBoost (Walker et al., 2022), while the combination of multiple images may involve stacking them as channels in a Convolutional Neural Network (CNN) framework (Taleb et al., 2021).

Early fusion combines raw data from different modalities at the input level, preserving the original information from each modality without potentially losing essential details. Moreover, it can simplify the model architecture by merging data before analysis, potentially reducing the computational complexity during the learning process. Disadvantages include challenges when combining data from multiple modalities at an early layer in the network, which can result in an imbalance of ‘data richness’ from each modality (Li et al., 2022). For example, with vision and language inputs, both the vision and the language data are passed through light feature extractors to convert the data to the desired feature space. However, the vision features require more processing by the machine learning model compared to the language features, primarily due to the

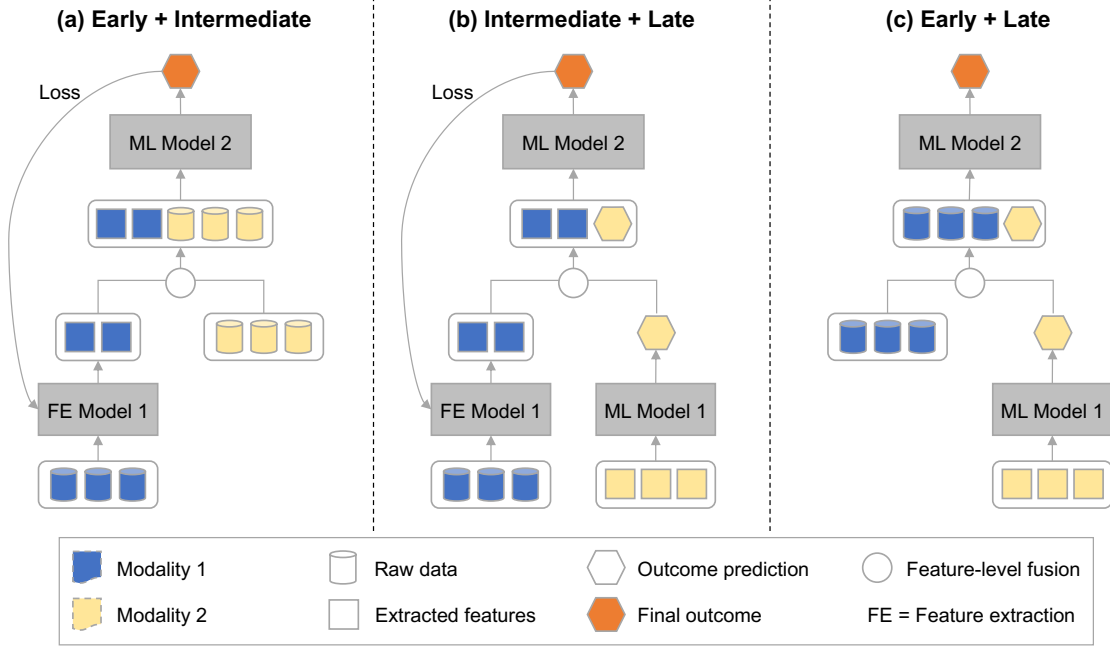


Figure 4: **Examples of mixed fusion architectures:** (a) The loss is only back-propagated for some modalities (blue) while others (yellow) are fused at a later step; (b) Similar to (a), but predictions from only one modality are used; (c) Features from one modality (blue) are combined with predictions from another modality (yellow).

data richness present in the image itself. Consequently, the machine learning model would allocate disproportionate attention to the vision modality. Moreover, the low-level features of a single modality may not necessarily provide any semantic meaning in a way that can be fused with another modalities. For instance, word embeddings (low-level features used for language data (Vaswani et al., 2017)) lack the capability to deduce information from edges in an image, which are typically the kind of features extracted in computer vision models such as AlexNet (Krizhevsky et al., 2012). This necessitates modality-specific feature encoding to standardise features in terms of semantic richness before forwarding them to a machine learning model for task-specific processing. Another challenge associated with early fusion is its potential inflexibility in accommodating new data sources or modalities without necessitating the retraining of the entire model.

Early fusion has been implemented in clinical machine learning across multiple settings, such as for predictive tasks in cardiological (Brugnara et al., 2020; Walker et al., 2022), oncological (Nie et al., 2019; Silva and Rohr, 2020; Vanguri et al., 2022) and neurological domains (Achalia et al., 2020; Li and Fan, 2019). In Duvieusart et al. (2022) an XGBoost

model was used to integrate laboratory results, demographic information, vital signs and image features for cardiomegaly prediction. Similarly, [Hyun et al. \(2019\)](#) fused features from CT and PET scans with demographic data for lung cancer diagnosis and [Dwivedi et al. \(2022\)](#) fused MRI and PET images with demographic and genetic data for Alzheimer’s disease prediction.

Intermediate fusion. In *intermediate fusion* ([Kline et al., 2022](#); [Ramachandram and Taylor, 2017](#)) (also referred to as *joint fusion* ([Huang et al., 2020b,c](#))), different data modalities are first processed by individual models, before the extracted features are combined and fed into a final prediction model (see Figure 3(b)). Unlike early fusion, here the loss function is back-propagated through the feature extraction model in order to generate improved feature representations during each iteration of training ([Huang et al., 2020b](#)). A common training approach for intermediate fusion involves first pre-training models separately on individual modalities ([Balakrishnan and Priya, 2018](#); [Duvieusart et al., 2022](#); [Palepu and Beam, 2022](#)). Subsequently, the weights of these models are frozen, their outputs concatenated and the final model is trained. During the final training step, certain weights may remain fixed for a specific duration or within specific modalities. However, it is crucial for a process to be considered intermediate fusion that at least some weights are unfrozen at least once during the training process ([Huang et al., 2020b](#)).

Advantages of intermediate fusion include the capability to extract and integrate features at various levels of abstraction, potentially capturing more complex interactions between modalities. Moreover, it provides flexibility in the model architecture, enabling separate processing pathways for different data types before fusion, preserving modality-specific information. However, intermediate fusion has its own complexities. The additional interactions and combinations can make the model complex and data-hungry ([Huang et al., 2020b](#)). Deep learning models, which are often used in intermediate fusion, typically require a large amount of data to effectively learn from the intricate feature interactions.

Intermediate fusion has been used in cancer prediction, where it is common practice to combine pathological images with additional demographic and genomics data (Cheerla and Gevaert, 2019; Li et al., 2020a; Schulz et al., 2021). Similarly, within cardiology intermediate fusion has been used to combine X-ray images with demographic information, biomarkers and clinical measurements (Baltruschat et al., 2019; Chauhan et al., 2020; Duvieusart et al., 2022). In addition, intermediate fusion has been used to predict brain disorders by combining MRI images with demographic information and other clinical and genetic information (El-Sappagh et al., 2020; Ghosal et al., 2021; Pölsterl et al., 2021; Spasov et al., 2018).

Late fusion. In *late fusion* (Huang et al., 2020b,c; Kline et al., 2022) (also referred to as *decision level fusion* (Ayesha et al., 2021; Behrad and Abadeh, 2022; Dolly and Nisa, 2019)), distinct models are run on separate data modalities (either raw or extracted features) and the resulting predictions are merged through an aggregation function or an auxiliary model (Baltrušaitis et al., 2018; Gao et al., 2020) (see Figure 3(c)).

One advantage of late fusion is that it can easily deal with missing data for patients. For example, CLIP (Contrastive Language-Image Pre-Training) a network pre-trained using image and text data, is able to perform as an image classifier even without existing text data being provided as an input during inference (zero-shot learning) (Radford et al., 2021). On the other hand, with late fusion it is not possible to model interactions and relationships between different modalities, which may lead to a loss of information (Gao et al., 2020; Huang et al., 2020a,b). Further, late fusion has its own set of challenges. The integration of results from different models can be complex and determining the optimal way of combining them is not always straightforward (Salekin et al., 2021).

Late fusion has been used in the diagnosis of cognitive impairment by combining MRI image predictions with demographic data and cognitive assessment scores (Qiu et al., 2018); in cancer prediction by merging MRI images with various biomarkers (Reda et al., 2018); and in COVID prediction by integrating CT scans with demographic information and clinical measurements (Zhou et al., 2021a).

Mixed fusion. The previously discussed fusion methods can be combined (or mixed) in a way such that each modality is processed in its most optimal way, mitigating the modality imbalances caused by early fusion, but also modelling the inter-modality dependencies which cannot be done using late fusion. For example, when processing vision-language models, it has been found that modelling data jointly works best when image data has been independently processed to a certain degree before combining with language data (Li et al., 2022; Xu et al., 2023a).

Whilst mixed fusion comes with the advantage of handling modality imbalances (as illustrated in Figure 4, which allows for tailored integration of data modalities), it is challenging to design these networks as deciding at which point into the processing pipeline the modalities need to be combined requires careful consideration (Xu et al., 2023b).

In healthcare, mixed fusion mostly occurs when one of the data modalities is an image. This is common for multi-modal histomics data processing, where histopathology images are processed up to the global feature level before combining with tabular genomic data (Chen et al., 2021b, 2022). Other studies used various fusion techniques for a specific disease diagnosis, combining features from MRI images with demographic and genetic data and clinical measurements (Venugopalan et al., 2021).

2.4.2 Feature-level fusion

For early and intermediate fusion, various methods for fusing features can be distinguished. While we only provide a brief summary (Figure 5), a more extended overview can be found in the literature (Baltrušaitis et al., 2018; Cui et al., 2022; Gao et al., 2020; Ramachandram and Taylor, 2017).

Concatenation. *Concatenation* is carried out when feature vectors are appended to form a single, longer vector (Cui et al., 2022) (see Figure 5(a)).

Since this approach is straightforward, flexible to different lengths of input and does not require additional parameters to tune, it is frequently used in practice (Holste

et al., 2021). However, it has its drawbacks: concatenating features creates long vectors, which may lead to overfitting, especially when training datasets are not sufficiently large (Duvieusart et al., 2022; Walker et al., 2022; Yan et al., 2021).

Concatenation was used in Duvieusart et al. (2022) to combine biomarkers extracted from images with demographic information, aggregated vital signs and lab results. Similarly, El-Sappagh et al. (2020) combined various features extracted from time-series data in Alzheimer’s disease progression detection.

Operation-based. In *operation-based* fusion, feature vectors are combined by element-wise operations, i.e. operations are performed on corresponding elements in two or more arrays or matrices of the same dimensions (see Figure 5(b)).

This approach requires vectors to have the same shape, using either element-wise or channel-wise multiplication. Element-wise multiplication in attention layers refers to multiplying corresponding elements of two matrices, whereas channel-wise multiplication involves multiplying entire channels (a specific dimension in a multi-dimensional array) of one matrix with another, treating the channel as a single entity. Alternatively, feature vectors can be combined in an attention-based manner (using attention layers) (Cui et al., 2022; Pölsterl et al., 2021; Schulz et al., 2021; Silva and Rohr, 2020), using one feature vector as attention weights for the other. Moreover, in tensor-based multiplication, feature vectors are combined by conducting outer products, with the aim of providing more information beyond that of the individual features (Braman et al., 2021; Chen et al., 2020; Cui et al., 2022).

Examples include Chen et al. (2020) where the authors used tensor-based fusion to correlate pathological images with genomics data to improve diagnostic accuracy; and Schulz et al. (2021) where attention layers were used for cancer prognosis prediction.

Learning-based. In *learning-based* fusion, feature vectors are combined using a machine learning model (see Figure 5(c)).

A specific example of learning-based methods is the Graph Convolutional Network

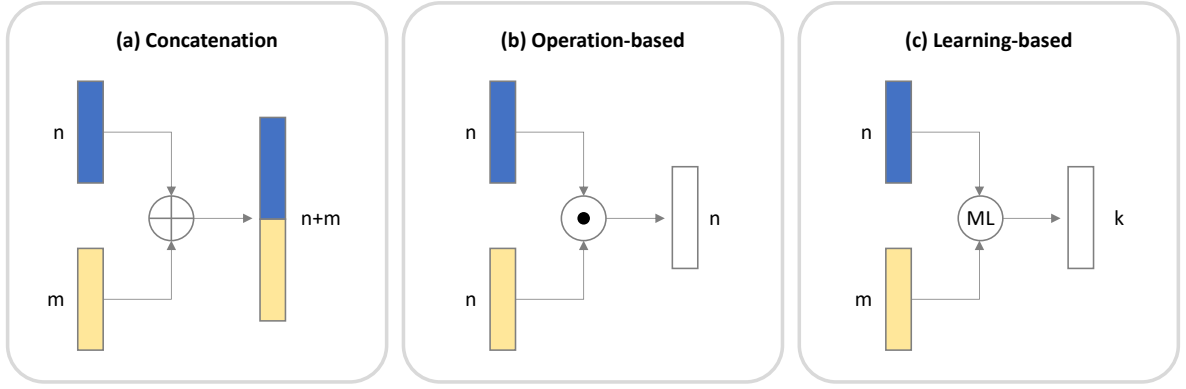


Figure 5: **Feature level fusion:** (a) Concatenation involves merging feature vectors end-to-end. (b) Operation-based methods combine vectors via element-wise mathematical operations or attention mechanisms, necessitating same-shaped vectors. (c) Learning-based fusion uses machine learning to reconstruct original features in a shared informative space.

(GCN), which employs nodes and edges to establish relationships between input data (e.g. nodes representing image features and edges signifying similarity between image and non-image features), ultimately learning a common feature vector (Cao et al., 2021; Cui et al., 2022; Parisot et al., 2018). Similarly, in image-based networks (e.g. CNNs, ViTs), different ‘image channels’ can be used to combine information from different modalities (Taleb et al., 2021).

In Pölsterl et al. (2021) the authors used an extension to conventional CNNs to combine 3D images with tabular data in Alzheimer’s disease prediction; and Parisot et al. (2018) used GCNs to combine MRI and clinical features in Alzheimer’s disease prediction.

2.5 Multimodal datasets and studies

2.5.1 Multimodal datasets

In Table 21 we provide an overview of various open-source, multimodal healthcare datasets and archives of datasets that contain multimodal information. We have primarily focused on multimodal content, which includes imaging data. These datasets contain demographic information and a subset contain extensive electronic health records comprising treatment histories, clinical metrics and in rare instances, full-text medical reports and vital sign time-series data.

Using multimodal datasets effectively requires a strategic approach that integrates diverse data types to enrich analysis and decision-making processes. Key considerations include the selection of additional modalities to address specific research questions, ensuring data quality and compatibility across types and adopting advanced analytic techniques capable of handling the heterogeneity inherent in multimodal data. Another important factor to consider is the allocation of sufficient computational resources to manage and analyse large and often complex, multimodal datasets efficiently.

2.5.2 Multimodal studies

We have conducted an overview of recent studies that employ multimodal approaches in healthcare. We categorise these studies into four groups: those predicting brain disorders (Table 22), those making cancer predictions (Table 23), those making predictions related to chest (Table 24) or skin conditions and other diseases (Table 25). Although our list is not exhaustive, we have focused on offering a comprehensive overview of noteworthy examples of multimodal deep learning applications in healthcare, organised by their respective application areas.

Most of the studies combined imaging data with tabular data (e.g. demographics, clinical measurements). The most commonly employed imaging modalities were MRI, followed by PET, CT and general X-ray imaging. Dermoscopic images were the predominant choice for skin disease studies. In some instances, studies incorporated additional assessment scores or markers extracted from images as tabular data alongside patient metadata. Only a limited number of studies integrated free text data from reports or time-series data, such as audio recordings or vital signs. Typically, these studies focused on feature extraction from time-series data rather than using the raw time-series. It is worth noting that many studies related to cancer and brain disorders used tabular features extracted from genomic data, evident in research areas including Alzheimer’s diagnosis ([Huang and Chung, 2020](#); [Venugopalan et al., 2021](#)) and breast cancer diagnosis ([Duanmu et al., 2020](#); [Liu and Hu, 2019](#)).

The predominant analysis method used on imaging data were CNNs. For time-series

data, simple feature extraction using aggregation methods, such as calculating the mean and standard deviation, were commonly applied. For analysing textual data in reports, studies typically used RNNs. There were various techniques used for tabular data, depending on the fusion technique applied. Early and intermediate fusion were the most frequently used approaches. Late fusion was less common and typically involved independently calculated prediction scores for each modality before they were combined. The stage at which features were combined varied significantly across studies. Transfer learning has become a common strategy, particularly in studies that encounter limited data (Walker et al., 2022). While the majority of studies employed supervised learning, interest in unsupervised and semi-supervised approaches due to the challenges of acquiring labelled data was noted (Chauhan et al., 2020; Huang and Chung, 2020; Parisot et al., 2018; Taleb et al., 2021).

2.6 Conclusions and future research

Having reviewed the relevant data modalities, model development strategies and fusion approaches in healthcare, we now discuss the potential directions for future research in the field.

2.6.1 Multimodal data availability, integration and resource requirements

One of the major challenges in developing multimodal machine learning approaches for healthcare lies in the limited availability of high quality multimodal datasets, comprising of diverse clinical modalities including clinical notes, images and time-series data, available for the same individual (Shaik et al., 2023). However, acquiring multimodal datasets for machine learning in healthcare poses significant challenges related to privacy regulations, ethical and regulatory complexities, limited collaborations and resource constraints (Greenhalgh et al., 2017; Rajpurkar et al., 2022). From the technical point of view the difficulties in aligning different modalities (often linked with longitudinal nature of data) complicates even further the creation of multimodal datasets suitable for the use with machine learning. Extensive data pre-processing can often help to improve

model performance (Krones et al., 2023). To overcome many of those obstacles would require the creation of clearer regulatory framework and more efficient collaboration between the healthcare and research communities.

Another challenge is that multimodal machine learning approaches often involve high computational complexity and significant resource requirements. The integration of diverse data types demands extensive preprocessing, feature extraction and model training, which can be computationally intensive. Models such as deep neural networks, particularly those employing intermediate or mixed fusion techniques, require substantial memory and processing power. This can lead to increased training times and the need for advanced hardware, posing challenges for practical implementation in resource-constrained environments.

2.6.2 Impact of emerging trends on multimodal AI in healthcare

Emerging trends that will impact multimodal AI in healthcare include federated learning, self-supervised learning, better data sources and system optimisations tailored to the unique demands of multimodal AI. Federated learning can improve multimodal AI by enabling models to learn from diverse, decentralised data without compromising patient privacy (Rieke et al., 2020), especially important when combining different healthcare data types. Self-supervised learning can be particularly helpful for multimodal AI, as it enables effective learning from vast unlabelled datasets, which often include complex, varied data types (Alayrac et al., 2020; Krishnan et al., 2022; Taleb et al., 2021). Specific performance demands highlight the need for system optimisations and hardware tailored to multimodal AI, addressing the unique computational and memory demands (Golden et al., 2024). And as discussed before, one of the key drivers for multimodal research will be the development of new multimodal datasets.

In addition, one of the biggest impacts could foundation models have, large, multipurpose models, typically pre-trained in a self-supervised manner on extensive datasets (Bommasani et al., 2021; Moor et al., 2023). They have demonstrated remarkable versatility in various domains including natural language processing (Devlin et al., 2018; OpenAI,

2023) and multimodal tasks (Wang et al., 2022a). However, their adoption in the medical domain is still in its early stages. This is partly due to the challenges of accessing diverse medical datasets and the complexity of the healthcare realm. Most of the current medical machine learning models were developed to address very specific tasks, limiting their adaptability and flexibility. Recent advances in foundation model research, such as multimodal architectures (Acosta et al., 2022) and self-supervised learning techniques (Krishnan et al., 2022), could potentially address these challenges. A particularly active trend is the combination of Large Language Model (LLM)s with other modalities, such as imaging, in vision-language foundation models that process image and text jointly (Acosta et al., 2022; Wang et al., 2022a). Such general-domain systems are increasingly used outside healthcare, and beyond raw predictive performance they can express the reasoning behind a prediction in natural language, which is attractive for clinical use. Recent evaluations even report general-purpose LLMs matching or exceeding specialised clinical tools on some medical benchmarks (Vishwanath et al., 2026). Foundation models are also promising precisely when datasets are small, a common problem in healthcare, because the general features needed for a downstream task are largely learned during large-scale pre-training and only need to be adapted through fine-tuning. We return to these directions, and to the question of whether task-specific models are still needed, as future work in [Chapter 7](#).

These developments may pave the way for generalist medical AI models, designed to handle a broad range of medical tasks and data types, including images, text and structured information (Moor et al., 2023; Tu et al., 2023). Generalist medical AI models aim to address healthcare challenges by offering adaptable and versatile solutions. These models have the potential to dynamically adapt to new tasks without extensive retraining (Alayrac et al., 2020; Brown et al., 2020), handle diverse data modalities and represent medical knowledge for advanced reasoning (Acosta et al., 2022). However, realising this potential requires addressing substantial challenges, such as refining model interpretability and ensuring ethical and responsible AI deployment. By overcoming these obstacles, future research could unlock the promise of generalist medical AI models

across various medical applications, potentially improving patient care and healthcare practices (Zhou et al., 2023c).

2.6.3 Deployment and clinical impact

Deploying multimodal AI in healthcare faces key challenges, particularly in integrating AI into diverse healthcare workflows (Kelly et al., 2019; Rajpurkar et al., 2022). Differences in patient management systems, clinical practices and digital infrastructures require adaptable and compatible systems. Regulatory hurdles are also amplified for multimodal AI due to the varied nature of data sources and the need for stringent privacy measures for each data type. Building trust in multimodal AI systems requires clear demonstration of their ability to handle diverse data securely and effectively. Addressing these specific challenges is critical for the successful deployment and widespread acceptance of multimodal AI in healthcare (Greenhalgh et al., 2017; Handley, 2023).

As noted earlier, interpretability is important in healthcare because it fosters trust and ensures the behaviour of the model adheres to ethical and legal guidelines. Implementing interpretability in multimodal approaches presents unique challenges due to the complexity and heterogeneity of combined data types (Joshi et al., 2021; Wang et al., 2021b). Aligning interpretability methods like saliency maps (Haghighi et al., 2022) for images and attention mechanisms (Vaswani et al., 2017) for text can be challenging. Additionally, ensuring that these multimodal models maintain transparency without compromising performance is complex, particularly when integrating diverse data sources in clinical settings, necessitating rigorous alignment with ethical and legal standards.

2.6.4 Conclusions

In our comprehensive review of multimodal deep learning fusion approaches within healthcare, we identified several key insights. Modality-level fusion techniques, which include early, intermediate and late fusion stages, have been shown potential in improving predictive accuracy across many healthcare areas, especially in disease diagnosis and prognosis. Our study of multimodal datasets revealed important resources that bring

together imaging data, demographic details, electronic health records and clinical measurements. It is noteworthy how disease prediction research has effectively used multimodal deep learning to achieve better results in diagnosis, treatment planning and survival rate predictions. The frequent use of MRI among imaging methods highlights its importance in the expansion of multimodal healthcare applications. Transfer learning has been crucial in overcoming challenges related to data scarcity. While supervised learning remains the primary approach, the popularity of semi-supervised and unsupervised methods is increasing. This reflects a growing interest in addressing healthcare issues where expert-annotated data are limited.

To address the current limitations of multimodal approaches, a comprehensive roadmap should include several key initiatives. Enhancing multimodal data availability requires the creation of extensive, standardised datasets that integrate various data types while ensuring patient privacy and compliance with regulations. Improved data integration techniques are necessary to harmonise diverse modalities, supported by advanced pre-processing and fusion methods. To mitigate resource requirements, the development of more efficient algorithms and leveraging cloud-based solutions can optimise computational demands. For deployment and clinical impact, establishing robust validation frameworks and pilot studies can facilitate the seamless integration of multimodal models into clinical workflows. Finally, improving explainability involves advancing XAI techniques and visualisation tools to ensure that model decisions are transparent and comprehensible to clinicians, thereby fostering trust and adherence to ethical standards.

2.6.5 From the review to the thesis contributions

The gap this thesis addresses. Taken together, the evidence above points to one central gap. Multimodal data are abundant in routine ICU care, yet there is no systematic, clinically aligned way to decide which fusion strategy to use, which modalities add value and when. Existing studies compare only a few approaches on a few modalities, rarely on data that reflect the intensive care setting, and rarely account for the staggered availability of clinical data or for subgroup robustness. Closing this gap requires both

the technical means to bring all relevant modalities into a single pipeline and a principled way to evaluate when their integration helps.

From the gap to the contributions. I addressed this gap as a design science research process (Hevner et al., 2004; Peffers et al., 2007), in which each of the following chapters builds and evaluates a concrete solution to one part of the problem, rather than presenting a set of unconnected papers. Figure 6 summarises this thread. The review (Chapter 2) defines the problem and the objectives. The first two solutions remove data-representation barriers: Chapter 3 makes paper ECGs machine-readable, and Chapter 4 integrates high-frequency waveforms without expert feature engineering. With these representations in place, Chapter 5 builds and evaluates the systematic fusion framework, and Chapter 6 adds the robustness and fairness layer. Each solution is demonstrated on publicly available data, namely the PhysioNet Challenges and MIMIC-IV, and tested against the research questions from Chapter 1. Throughout, the work was conducted in close collaboration with ICU clinicians, who shaped the choice of conditions, the labels and features and the interpretation of the results, so that clinical input is part of the methodology rather than an afterthought. The five resulting papers (Section 1.4) form the communication step of this approach.

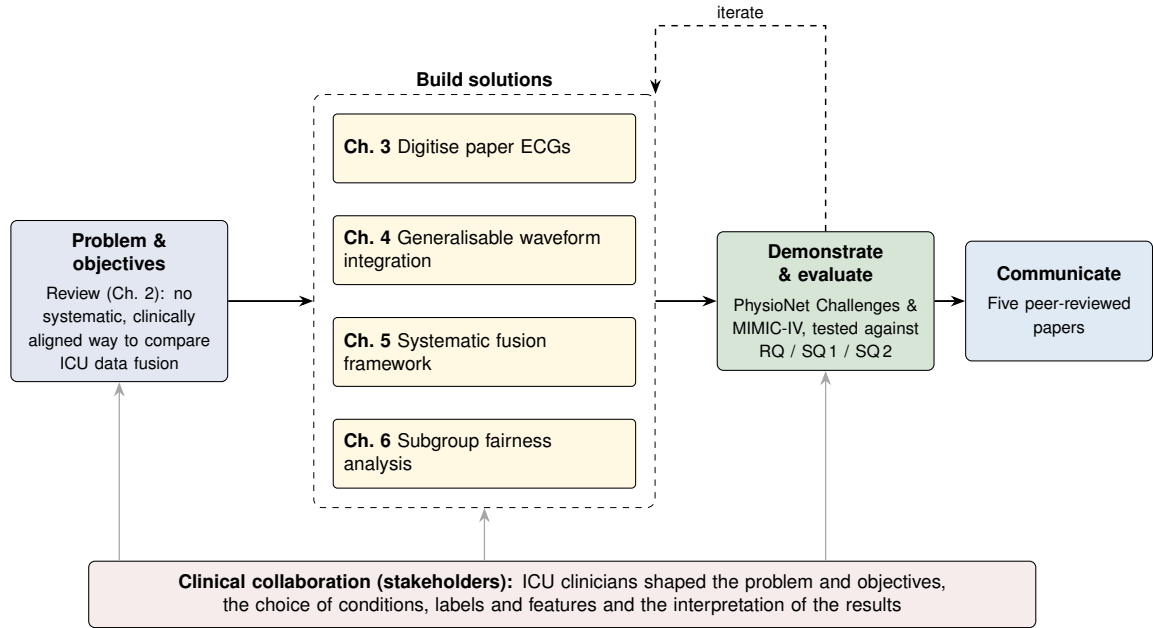


Figure 6: **Design science research thread of the thesis.** The review identifies the problem and objectives; the four empirical chapters build solutions, namely digitisation ([Chapter 3](#)), generalisable waveform integration ([Chapter 4](#)), a systematic fusion framework ([Chapter 5](#)) and subgroup fairness ([Chapter 6](#)); these are demonstrated and evaluated on the PhysioNet Challenges and MIMIC-IV against the research questions, and communicated through five peer-reviewed papers. Throughout, ICU clinicians act as stakeholders who shaped the problem, the data choices and the interpretation of the results. The dashed arrow indicates the iterative nature of the process.

3 Digitising ECG waveforms for use in multimodal pipelines

Extended chapter based on the paper ([Krones et al., 2024](#))*:

Krones F, Walker B, Lyons T, Mahdi A. *Combining Hough Transform and Deep Learning Approaches to Reconstruct ECG Signals From Printouts*. **Computing in Cardiology** (2024).

*Winner of the 2024 PhysioNet Challenge

 [felixkrones/physionet24](#)

Abstract. This chapter presents our team’s winning contribution to the 2024 George B. Moody PhysioNet Challenge on reconstructing waveform signals (ECGs) from printouts. Although many ECGs are now recorded digitally, paper ECGs remain common worldwide—either for transferring records between clinicians, adding annotations, or due to limited access to digital devices. Digitising them helps to build more diverse datasets and to enable automated analyses. However, the presence of varying recording standards and poor image quality requires a data-centric approach for developing robust models that can generalise effectively. Our approach combines the creation of a diverse training set, the rotation of images using the Hough transform, a U-Net based segmentation model to identify individual signals and mask vectorisation to reconstruct the signals. We assessed the performance of our models using the 10-fold stratified cross-validation (CV) split of 21,799 recordings proposed by the PTB-XL dataset. The study shows the challenges of building robust, generalisable, digitisation approaches. Such models require large amounts of resources (data, time and computational power) but have great potential in diversifying the data available.

Contents

3.1	Introduction	55
3.2	Related work	56
3.3	Methodology	66
3.4	Results	72
3.5	Discussion	73

3.1 Introduction

This chapter develops the first building block for the overarching question posed in [Chapter 1](#), namely under which conditions multimodal integration improves clinical prediction. The review in [Chapter 2](#) identified the scarcity of large, diverse multimodal datasets as a central barrier, and waveforms such as ECGs are often locked in paper records. Reliable digitisation is therefore a prerequisite for the fusion experiments in [Chapter 5](#) and [Chapter 6](#).

Cardiovascular diseases are the leading global cause of mortality ([World Health Organisation, 2021](#)) and Electrocardiograms (ECGs) play a central role in their diagnosis and management. Since the first practical ECG device developed by Willem Einthoven in 1895, the technology has undergone substantial evolution—from early mechanical systems to modern digital solutions. While the development of digital ECG devices has substantially improved the accessibility and automation of cardiac diagnostics, paper-based ECGs remain widespread, particularly in regions with limited resources ([Reyna et al., 2024b](#)).

Printed ECGs, which became widespread as early as the mid-20th century, continue to be used due to their portability, ease of annotation and compatibility with legacy systems. Although many newly recorded ECGs are now captured digitally, it is estimated that billions of historical recordings exist only in printed form, especially in the Global South ([Reyna et al., 2024b](#); [Shivashankara et al., 2024](#)). These paper ECGs are a rich resource, encompassing decades of clinical data across diverse populations and healthcare environments. Unlocking their diagnostic potential requires reliable digitisation methods.

Digitising ECGs offers several advantages: it allows the integration of legacy data into modern datasets, supports the development of more representative diagnostic models and lowers the barriers imposed by proprietary digital systems. However, this process is challenging due to variations in paper quality, scanning resolution and graphical layout ([Baydoun et al., 2019](#); [Fortune et al., 2021](#); [Li et al., 2020b](#); [Wu et al., 2022](#)).

Improvement therefore requires models that are capable of recovering signal information from variable-quality scans and heterogeneous print standards.

This has led to the following research question:

RQ: *How can waveform signals be reconstructed from paper printouts using a generalisable digitisation model across diverse layouts and image qualities?*

In response to mentioned challenges, we focused on the digitisation of ECG signals from paper printouts, as part of the PhysioNet Challenge 2024. We developed a flexible and robust pipeline capable of addressing the variability present in real-world ECG images. The approach integrates traditional image processing techniques with deep learning methods: image rotation was corrected using the Hough line transform, waveform regions were segmented using a deep learning model and the resulting masks were vectorised to reconstruct the underlying signals. To improve generalisability across diverse layouts and visual styles, we created multiple image representations for each recording and conducted extensive cross-validation on 21,799 ECGs from the PTB-XL dataset ([Wagner et al., 2020](#)). This framework was designed to support accurate and reliable digitisation under heterogeneous recording conditions, contributing towards more inclusive access to digital ECG data. We compared our approach with the other top-performing Challenge submissions and showed that it achieved more than twice the score of the next best model.

The core contribution of this chapter: we developed a generalisable pipeline combining image processing and deep learning to reconstruct waveform signals from paper printouts across diverse layouts and qualities.

3.2 Related work

ECGs are widely used in cardiovascular diagnostics and play an important role in assessing cardiac condition. They capture both the electrical activity of the heart and its structural features, such as the arrangement of chambers and valves, thus providing

a comprehensive evaluation of cardiac health (Tison et al., 2019). Hence, the interest in integrating ECG data into multimodal prediction models continues to grow and the need for robust digitisation methods becomes increasingly evident. This is particularly relevant because, despite the widespread adoption of digital recording technologies, paper-based ECGs are still commonly used (Reyna et al., 2024b). Converting paper ECG records into digital format not only increases the volume of available data but also supports automated analyses and enables multimodal approaches that may improve predictive performance.

3.2.1 Clinical relevance

Automated ECG analysis has demonstrated substantial potential in clinical decision support, with deep learning models accurately predicting a broad spectrum of both cardiac and non-cardiac conditions based solely on ECG data (Strodthoff et al., 2024). Integrating ECG waveforms into multimodal pipelines can further improve predictive performance. For example, combining ECG with other modalities, such as chest X-rays, demographic information and laboratory results, has been shown to outperform traditional decision tree algorithms in localising accessory pathways in Wolff–Parkinson–White syndrome (Nishimori et al., 2021). Similarly, in emergency care settings, multimodal approaches combining ECG waveforms with clinical data have shown high accuracy in predicting diagnoses and patient deterioration (Lopez Alcaraz and Strodthoff, 2024). Furthermore, it has been shown that models using raw ECG waveform data can outperform models based solely on structured or feature-extracted data (Alcaraz et al., 2025). Integrating waveform data with patient-level features such as vital signs and laboratory trends allows for a more complete understanding of patient status and enhances decision support systems (Lopez Alcaraz and Strodthoff, 2024).

3.2.2 ECG digitisation methods

The increasing interest in multimodal predictive models has highlighted the importance of reliable ECG digitisation. Although digital ECG recorders are now widely available, the

use of paper ECGs remains common and large archives of paper ECGs are increasingly recognised as a valuable resource for retrospective research and AI development (Reyna et al., 2024b). Previous work showed that converting printed ECGs to digital signals is technically feasible but methodologically challenging as prior digitisation efforts were often constrained by limited data diversity, which reduced model robustness and generalisability (Lence et al., 2023; Waits and Soliman, 2017).

Early work focused on converting scanned ECG plots into digital time series using basic image processing. For example, Widman and Hines (1991) introduced a semi-automated method in which paper ECGs from 8 patients were scanned using a common desktop scanner, an operator manually edited each image to remove grid lines and stray marks and a line-finding algorithm was used to digitise the remaining signal into a digital waveform. Similarly, Lawson et al. (1995) developed a system for standard 12-lead ECGs in which paper recordings are scanned, the operator manually calibrates the greyscale threshold to separate the signal from surface noise and grid lines and an automatic algorithm then follows the darkest path of the trace; the user can add or delete points where the algorithm is wrong. In total, the authors analysed 84 leads. Wang and Mital (1996) proposed a stepwise approach in which different pre-filtering steps were first applied to remove grid lines and noise, based on heuristic line-thickness assumptions. Next, a thinning algorithm was used to identify the signal line, followed by post-filtering to remove additional noise.

Subsequent methods in the 2000s introduced more automated contour-tracking and grid-based approaches. For example, Mitra et al. (2004) proposed a pipeline in which single-lead waveforms printed on paper are scanned with a flatbed scanner, converted to a binary image and the grid is removed using thresholding; a thinning step then reduces the trace to a single-pixel centre line whose coordinates are extracted and ordered to reconstruct a one-dimensional time series. Badilini et al. (2005) developed ECGScan for standard 12-lead ECGs, where the user first draws rectangles around each lead and can optionally place anchor points that the digitised trace must pass through, while the software automatically detects the ECG grid, estimates and corrects

the page rotation and uses an contour model to follow the waveform and reconstruct a continuous signal; they evaluated the system on 60 ECGs. In the 2010s, [Ravichandran et al. \(2013\)](#) introduced a MATLAB-based tool in which the ECG region is segmented and the user marks two points on a grid line for skew correction, after which the algorithm automatically detects the grid, corrects the page rotation, converts the image to black-and-white, applies filtering and interpolation to suppress the noise introduced by grid removal and finally recovers the waveform by scanning each column of the image. Other rule-based methods, surveyed in narrative and systematic reviews ([Lence et al., 2023](#); [Waits and Soliman, 2017](#)), typically follow the same three stages of straightening the page, removing the grid and thresholding the image and then extracting the signal either by scanning columns, grouping connected pixels or finding the most likely path of the trace, but they still depend on relatively standardised layouts, often retain at least some user interaction and have mostly been evaluated on small, curated collections of ECGs, which limits comparability and generalisability.

More recent work has focused on scalable, semi- or fully-automated digitisation, but many approaches still rely on pixel thresholding and handcrafted pre-processing, using a small set of images. For instance, [Baydoun et al. \(2019\)](#) developed a MATLAB-based algorithm that segments scanned ECG images into regions of interest using a sliding mask and comparing pixel standard deviations, followed by pixel thresholding to remove background noise and grid lines, but applied it to only 30 scanned ECG curves. Similarly, [Fortune et al. \(2021\)](#) introduced an open-source Python tool using pixel thresholding to segment signals, supplemented by manual pre-processing, including image rotation and localisation of the ECG leads. [Wu et al. \(2022\)](#) first redacted specific parts of the image, identified and removed red grid lines and then segmented the signals. These methods represent an evolution towards end-to-end automation, yet they remain sensitive to image quality and layout assumptions and are often validated on limited sets of ECG formats.

Deep learning has recently been adopted to improve robustness and reduce reliance on handcrafted rules. [Li et al. \(2020b\)](#) formulated ECG digitisation as an image

segmentation task and trained a U-Net-based architecture on grey-scale patches extracted from full ECG scans; the model predicts a binary mask of the ECG line, which is then converted to a one-dimensional signal using a previously published vectorisation procedure. [Mishra et al. \(2021\)](#) proposed a pipeline that assumes single-lead ECG images and uses a neural network to predict the binarisation threshold, followed by classical image processing to extract the ECG trace. In both approaches, however, the learned models operate only on standardised inputs (full-scan patches or single-lead images), so page-level tasks such as handling multi-lead layouts and integrating results into a fully automated workflow remain manual or heuristic. [Wu et al. \(2022\)](#) also use deep learning components in their fully automated algorithm, but still rely on extensive pre-processing steps such as colour-based grid removal and heuristic masking of non-signal areas.

Most recently, [Demolder et al. \(2025\)](#) presented a fully automated deep learning-based solution validated on a large image dataset with 6000 ECGs. Their system first standardises the image by detecting grid intersection points with a U-Net-like model, correcting geometric distortions so that the grid becomes regular again and resampling the image to this regular grid. A second U-Net then segments all ECG traces in the image. Post-processing assigns each trace to a lead based on the detected layout and converts pixel positions into calibrated time and voltage values. This approach shows that deep learning can achieve high precision and robustness across diverse acquisition conditions, provided sufficiently large and diverse training data. Nevertheless, [Lence et al. \(2023\)](#) and [Demolder et al. \(2025\)](#) both stress that public benchmarks and standardised evaluation metrics remain limited, complicating direct comparison and external validation of different algorithms.

In summary, prior digitisation methods range from early pixel thresholding and line following algorithms applied to small numbers of images, through semi-automated contour- and grid-based approaches, to recent deep learning models that automate more parts of the processing pipeline ([Table 3](#)). Many methods are reliant on specific image characteristics or manual pre-processing, reducing their generalisability. In contrast, our approach applies a deep learning network to the entire image, with pre-processing

Table 3: **Summary of ECG digitisation methods.** Representative approaches for converting paper-based or scanned ECGs into digital signals are summarised, ranging from classical image processing pipelines to more recent deep learning methods. The table illustrates the shift from rule-based grid detection and trace following towards data-driven segmentation and end-to-end learning, highlighting common steps such as grid removal, trace isolation and time–voltage calibration.

Reference	Year	Type	Key steps
Widman and Hines (1991)	1991	Classical	Scan paper ECGs; cleaned images are processed by a line-finding algorithm that traces the ECG and outputs a digital waveform.
Lawson et al. (1995)	1995	Classical	Scan 12-lead ECGs; apply a greyscale threshold; an algorithm follows the darkest path of the trace to reconstruct each lead.
Wang and Mital (1996)	1996	Classical	Pre-filtering to suppress grid and noise; thinning to obtain a single-pixel line; post-filtering to connect and smooth the trace.
Mitra et al. (2004)	2004	Classical	Scan single-lead ECGs; convert to binary; remove background and grid; thin the trace to a centre line and read its coordinates as a time series.
Badilini et al. (2005)	2005	Classical	Detect grid and page rotation; within lead boxes, an active contour follows the waveform and reconstructed signals are exported digitally.
Ravichandran et al. (2013)	2013	Classical	Segment ECG region; detect grid and correct skew; convert to black-and-white; filter artefacts; recover waveform by column-wise scanning.
Baydoun et al. (2019)	2019	Classical	Sliding mask and pixel variance to detect ECG regions; thresholding to remove background and grid; waveform extraction in ROIs.
Li et al. (2020b)	2020	Deep learning	U-Net-like network on greyscale patches to segment the ECG line; vectorisation converts mask to a one-dimensional signal.
Fortune et al. (2021)	2021	Classical	Scan ECG pages; remove grid and background; isolate the trace using connected components and morphology; map pixels to amplitude–time coordinates.
Mishra et al. (2021)	2021	Deep learning	Neural network predicts image-specific binarisation threshold; classical processing cleans the image and traces the waveform.
Wu et al. (2022)	2022	Deep learning	Combine layout and text detection, grid removal and signal extraction modules to segment and digitise all leads from scanned 12-lead ECGs.
Demolder et al. (2025)	2025	Deep learning	First U-Net-like model detects grid intersections and normalises geometry; second U-Net segments traces; post-processing assigns leads and calibrates time and voltage.

limited only to rotation correction, which was also fully automated. This design offers two key advantages: (a) it significantly improves performance compared to previous methods for ECG digitisation and (b) it can be readily applied to diverse data sources.

3.2.3 Hough transform

As discussed in the previous subsection, many digitisation pipelines rely on detecting straight structures in the image, such as grid lines or lead reference lines, to support reliable preprocessing before the actual waveform extraction. Among the available methods for identifying such linear elements, the Hough transform has become one of the most widely used because it can robustly detect straight lines even when images contain noise, faint print, or partial occlusions. Hough originally proposed that his transform can be used for the detection of straight-line segments in photographs of bubble chamber experiments, where charged particles leave visible tracks in a medium (Hough, 1962). Duda and Hart (1972) improved the algorithm by introducing a polar coordinate parametrisation using angle and radius, which helped the detection of vertical lines.

The Hough transform has been used in previous attempts to digitise ECG waveforms primarily to identify ECG baselines (Baydoun et al., 2019; Wu et al., 2022). In our work, its purpose lies specifically in detecting the underlying grid lines of the paper ECG, enabling the automated estimation of the rotation of the page prior to segmentation and vectorisation.

After an edge detector (see below) has been applied to an image, resulting in a set of edge points, one challenge is to identify geometrical shapes, including straight lines, from these points. The Hough transform addresses this by detecting lines that pass through multiple points in the image. Specifically, the transform aims to identify the lines that intersect the highest number of edge points or, at the very least, exceed a defined threshold.

Rather than working in the Cartesian coordinate system, the Hough transform operates in the polar coordinate system. Here, lines are represented by the intersections of multiple curves. Each curve corresponds to the family of all possible lines passing through a particular point in the image. The greater the number of intersecting curves, the more edge points lie on the corresponding line. This property enables a threshold to be set on the minimum number of intersections required to define a valid line.

In this framework, a point in the Cartesian coordinate system (x, y) is mapped to a curve in the polar coordinate system. Conversely, a point in the polar system (θ, r) represents a line in the Cartesian system. This duality is a central concept of the Hough transform.

The transformation proceeds as follows. For each point (x, y) in the Cartesian coordinate system, we consider the family of lines passing through it. In the polar coordinate system, these lines can be expressed as

$$r = x \cos(\theta) + y \sin(\theta).$$

For each (x, y) pair, we compute r for a range of θ values, generating a curve in the

polar system representing all lines passing through that point.

Next, we search for peaks in the polar parameter space, i.e., points (θ, r) with the highest number of curve intersections. These peaks correspond to lines in the image that pass through many edge points. To convert a peak (θ, r) back to a line in the Cartesian system, we use

$$y = -\frac{\cos(\theta)}{\sin(\theta)}x + \frac{r}{\sin(\theta)}.$$

The only question that remains is why $r = x \cos(\theta) + y \sin(\theta)$ represents the family of all lines passing through (x, y) . A straight line in Cartesian coordinates is typically expressed as $y = mx + c$, where m is the slope and c is the y-intercept and the line is drawn by computing y for various x . But this form becomes problematic for vertical lines, as the slope approaches infinity. In the polar coordinate system, this relationship can be written as

$$y = -\frac{\cos(\theta)}{\sin(\theta)}x + \frac{r}{\sin(\theta)},$$

which can be rearranged to

$$r = x \cos(\theta) + y \sin(\theta).$$

This equation describes all lines, parameterised by θ , passing through the point (x, y) in the Cartesian system.

3.2.4 Edge detection

As mentioned above, the Hough transform assumes the presence of a set of points of interest in the image. To identify such points, an edge detection algorithm is first applied. Edge detection is a fundamental step in image processing that aims to identify boundaries within an image where the intensity of pixel values changes abruptly. These boundaries often correspond to meaningful structural information, such as object contours, shadows, or texture changes and serve as critical input for higher-level tasks such as shape detection, segmentation, or image alignment.

The basic concept of edge detection is to analyse how quickly the brightness values

(pixel intensities) change across the image. This is commonly achieved by sliding a small matrix, called a kernel or filter, over the image. At each position, the kernel compares the surrounding pixel values to estimate how strongly and in which direction the intensity changes (i.e. it approximates how steep the change is in brightness). This process, often referred to as convolution (i.e. applying the kernel at each pixel location and calculating a weighted sum of the neighbouring values), results in a new image where high values mark locations with sharp intensity transitions. Such regions are likely to correspond to edges.

Let us assume a part of an image consisting of 4×4 pixels with the following greyscale intensity values:

$$I = \begin{bmatrix} 52 & 55 & 61 & 59 \\ 54 & 60 & 65 & 63 \\ 58 & 62 & 69 & 68 \\ 57 & 61 & 66 & 64 \end{bmatrix}$$

To detect edges, we can place a 3×3 kernel over the image. For instance, consider the Sobel kernel ([Sobel, 2014](#)) used to detect horizontal intensity changes:

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

We apply this kernel to the upper-left 3×3 region of the image:

$$I_{3 \times 3} = \begin{bmatrix} 52 & 55 & 61 \\ 54 & 60 & 65 \\ 58 & 62 & 69 \end{bmatrix}$$

To compute the response at the centre pixel, we multiply each element of the kernel with the corresponding pixel value and sum the results:

$$\begin{aligned}
R &= (-1) \cdot 52 + 0 \cdot 55 + 1 \cdot 61 \\
&\quad + (-2) \cdot 54 + 0 \cdot 60 + 2 \cdot 65 \\
&\quad + (-1) \cdot 58 + 0 \cdot 62 + 1 \cdot 69 \\
&= -52 + 0 + 61 - 108 + 0 + 130 - 58 + 0 + 69 \\
&= 42
\end{aligned}$$

The result $R = 42$ represents the strength of the horizontal intensity change at the centre pixel of that region, which we call G_x . A high positive or negative value indicates a strong edge; values close to zero suggest no substantial change in that direction. By repeating this process across the image, we obtain a new image that highlights horizontal edges. A similar computation using a vertical Sobel kernel detects vertical edges G_y . The two outputs can be combined to estimate the overall edge strength and direction. The Canny detector ([Canny, 1986](#)) extends gradient-based methods such as the Sobel operator with additional steps to improve edge localisation and reduce noise-induced artefacts. It performs two main tasks. First, it identifies the central line of an edge, thereby suppressing blurred border regions. Second, it removes non-dominant edges, which are likely to be noise.

The first step, locating clear, central edge points, consists of multiple sub-steps. Initially, (1) a Gaussian filter is applied to the image to reduce noise. This filter is a matrix that passes over the image and is characterised by a normal distribution with the highest values concentrated in the centre (typically in a 5×5 kernel). As a result, the image is smoothed, with more weight assigned to the central pixel. The parameter σ , which defines the Gaussian distribution, is another configurable input. Then, (2) Sobel operators are applied, yielding gradient components G_x and G_y . Next, (3) the magnitude and direction of the intensity gradients are computed (see below). Gradient directions are rounded to horizontal (0°), vertical (90°), or one of two diagonal directions

(45°, 135°). Once magnitude and direction are determined, (4) each pixel is examined along the direction of its gradient to determine whether it is a local maximum, thus indicating the centre of an edge. This step is often called *non-maximum suppression*.

The gradient magnitude and direction are calculated as follows: after obtaining G_x and G_y , the magnitude is given by

$$G = \sqrt{G_x^2 + G_y^2},$$

and the direction is computed using the two-argument arctangent function:

$$\theta = \tan^{-1} \left(\frac{G_y}{G_x} \right).$$

This angle θ is then quantised to one of the four discrete directions used in non-maximum suppression.

The second step of the algorithm is relatively straightforward: for edge selection, Canny applies two thresholds (often called *double thresholding*). Edges with gradient magnitudes above the upper threshold are retained automatically; those below the lower threshold are discarded. Pixels with values between the two thresholds are retained only if they are connected to a strong edge, a process known as *hysteresis thresholding*. These thresholds are typically parameters of the edge detection function. In our case, we used the `opencv` implementation, which also allows adjustment of the Sobel operator size.

3.3 Methodology

3.3.1 Data

The PTB-XL dataset (Goldberger et al., 2024; Wagner et al., 2020) contains 21,799 12-lead ECG recordings from 18,869 patients. The data was collected from October 1989 to June 1996 and publicly released in 2019. For each of the recordings, we created at least four images using randomly applied augmentation options from the `ecg-image-kit` (Deepanshi et al., 2024; Shivashankara et al., 2024). We used a maximum rotation of 30 degrees and default settings for the other augmentation options. Figure 7 provides

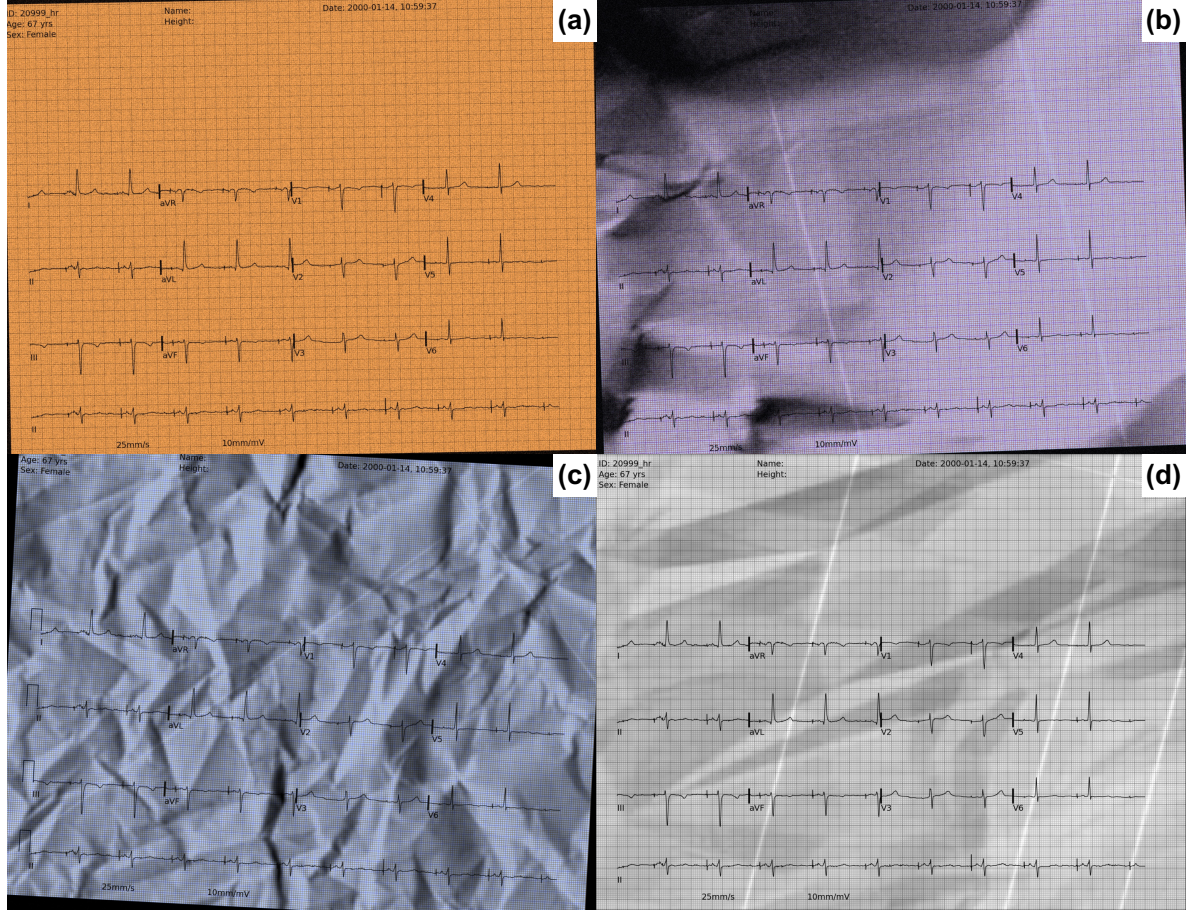


Figure 7: **Four example images of signal 20999:** (a) rotated by one degree, (b) rotated by two degrees with wrinkles and shadows, (c) rotated by three degrees with wrinkles and shadows, (d) not rotated with wrinkles and shadows.

examples of a single paper ECG trace that has undergone various degrees of rotation, often accompanied by wrinkles and shadows on the paper. [Table 4](#) summarises the various ways augmentation was applied during image generation. We kept the following variables identical across all images: signal order, relative signal positioning, grid scale and the number of full-length signals.

3.3.2 Model architecture

Our approach consisted of three major steps: image rotation, segmenting the image into individual signals and vectorising the segmented pixels. [Figure 8](#) shows a schematic diagram outlining this approach.

Rotation. We rotated the ECG images to standardise the subsequent tasks. To achieve

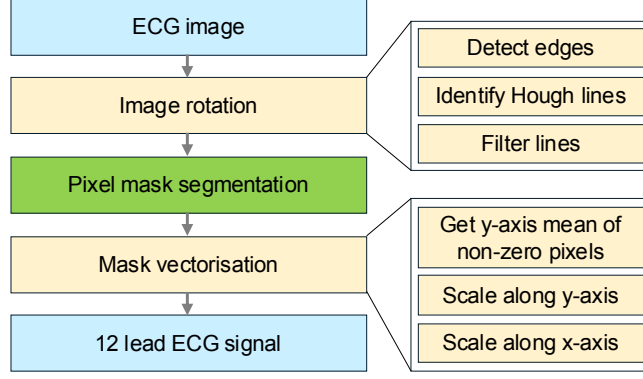


Figure 8: **Schematic overview of the ECG digitisation pipeline.** The workflow converts an ECG image into a 12-lead digital signal through a sequence of processing steps. *Blue* boxes denote data representations, *yellow* boxes indicate rule-based pre-processing and feature extraction, and the *green* box represents the trained deep learning segmentation model. Rule-based steps are used for image alignment and signal calibration, while learning-based segmentation enables robust trace extraction.

this, we used the `HoughLines()` method from the `opencv (cv2)` Python package. This method detects straight lines in an image by mapping edge points onto a polar coordinate system and finding intersections that indicate the presence of a line. We applied filters to these detected lines, retaining only those within a specific angle range and ensuring a minimum number of parallel lines. Using the angles of these filtered lines, we calculated the rotation to align the ECG waveforms horizontally.

Segmentation. For the segmentation we used nnU-Net (Isensee et al., 2021), which was developed by the German Cancer Research Center. nnU-Net automatically configures the image processing and network architecture based on the data, using three types of parameters. First, it uses a set of fixed parameters (which remain constant due to their robust nature), such as the loss function, primary data augmentation strategies and learning rate. Second, rule-based parameters are adapted to the dataset using predefined heuristic rules, adjusting for network topology, batch size and other factors. Finally, empirical parameters are determined through trial-and-error, during which elements such as the best U-net configuration for the dataset and the post-processing strategy are refined.

Figure 9 shows a segmentation example. To create the masks, we used two sets of pixels: the set provided by the `ecg-image-kit` (referred to as ‘sparse’) and an interpolated

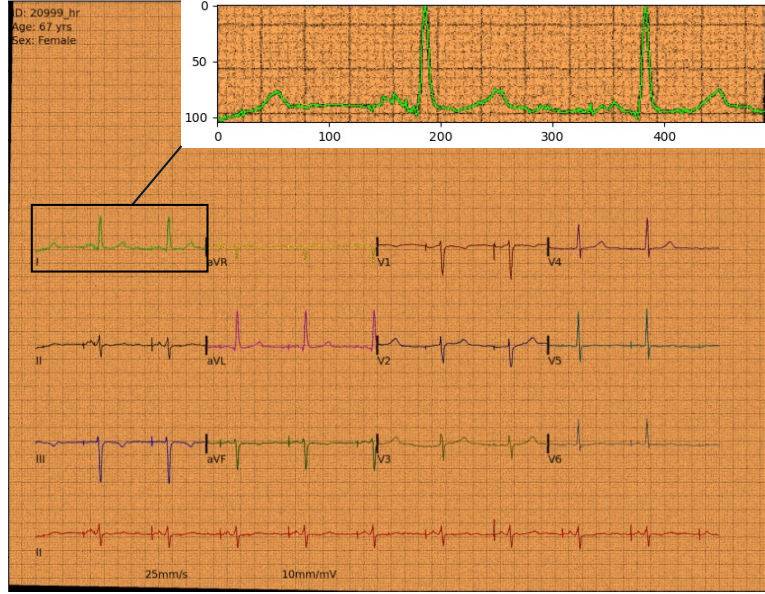


Figure 9: **Example of predicted masks for signal 20999.** The segmentation model segments the individual signals with labels (shown here through different colours).

version with increased pixel density (referred to as ‘dense’). Since the masks stored by default in `ecg-image-kit` are sparse, containing only a few points along the signal, we increased the pixel density by re-reading the stored JSON files containing the lead plots and applying linear interpolation between adjacent points. This interpolation added additional pixels, resulting in smoother and more detailed masks. The interpolated pixel arrays were saved back into the JSON files. Two different ground-truth representations were therefore used at the two stages of the pipeline. The segmentation model was trained against a pixel-by-pixel label mask, that is, a two-dimensional image of the same size as the input in which every pixel that belongs to a given lead trace carries the class index of that lead (1 to 12, see [Section 3.3.2](#)) and all remaining pixels are background. The final digitisation output and its evaluation, by contrast, used the original numeric signal, that is, the one-dimensional voltage array per lead as stored in PTB-XL, against which the vectorised prediction was compared using the SNR ([Equation 1](#)).

Vectorisation. In order to convert the mask, which is a matrix of pixel values, into a signal, the average position along the y-axis of the non-zero pixels in each column was used as signal value. Next, the signal was scaled to the correct amplitude using the grid size and resolution, which we assumed to be constant. Finally, the starting time of the

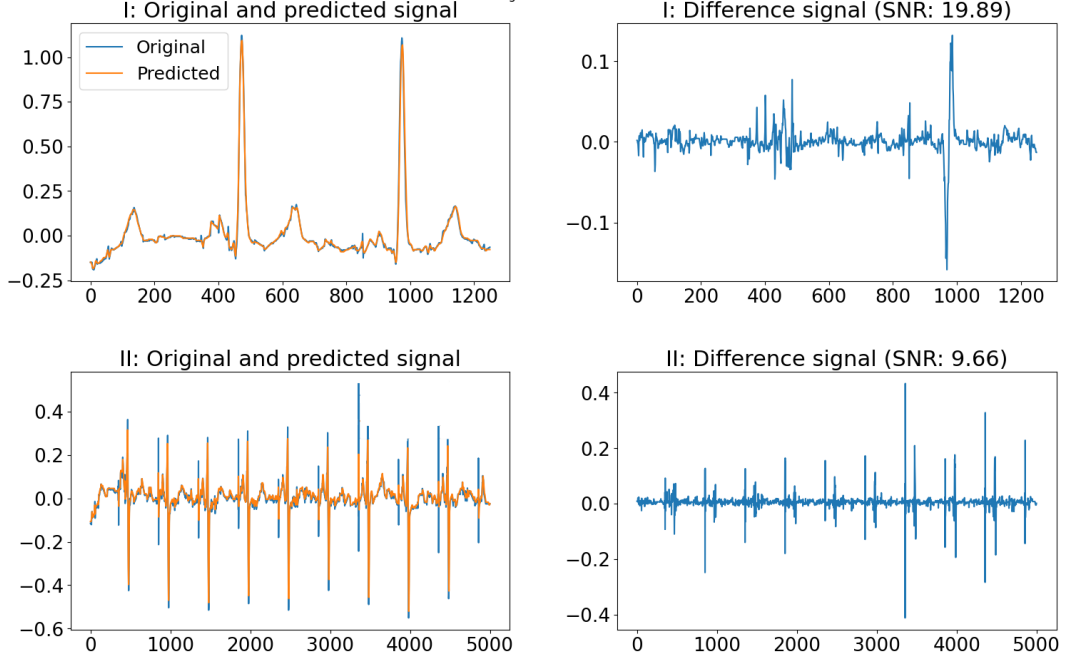


Figure 10: **Vectorisation example for signal 20999.** Shown is the vectorisation for two leads (I on top and II below) of signal 20999. *Left:* Original and predicted signal overlaid. *Right:* Difference between predicted and original signal.

signal was determined based on the position of the signal on the ECG. [Figure 10](#) shows an example of two leads from an ECG and the predicted signals from vectorising a mask of the signals.

3.3.3 Scoring

The metric used in the digitisation task was the signal-to-noise ratio (SNR) between the true signal y and predicted signal $\hat{y}$, which can be calculated using the formula:

$$\text{SNR} = 10 \cdot \log_{10} \frac{\sum_i y_i^2}{\sum_i (\hat{y}_i - y_i)^2}. \quad (1)$$

Letting the prediction be $\hat{y}_i = 0$ for all i provides a surprisingly strong baseline of $\text{SNR} = 0$ for this Challenge. This is due to the ECG signals having high and narrow peaks.

To evaluate our models, we used the 10-fold splits suggested in [Wagner et al. \(2020\)](#) due to their stratified sampling approach.

3.3.4 Statistical analysis

The term significant denotes results with a p-value of less than 0.05. All statistical tests were two-sided. Continuous clinical variables were compared between independent outcome groups using Welch’s two-sample t-test, which does not assume homogeneity of variances. To assess performance differences between models evaluated on identical cross-validation folds, we used paired-samples t-tests on fold-level performance metrics, thereby accounting for the dependence induced by shared data splits.

3.3.5 Benchmark comparisons

We compared our model to four other models that achieved an SNR score of at least 2. [Yoon et al. \(2024\)](#) (BAPORLab) trained a U-Net segmentation model on a three-channel image, where one channel represented the waveform masks, another the grid lines and the third the text in the image. The segmented masks were then digitised by first identifying short and long leads. The signal range was calculated using the grid line channel. [Verlyck et al. \(2024\)](#) introduced WAVIE, a three-step approach. First, they rotated the image using a neural network to detect image rotation. They then used a YOLO network ([Redmon et al., 2016](#)) for bounding box detection and nnU-Net ([Isensee et al., 2021](#)) for signal segmentation. To digitise the signals, they first applied a sliding window approach and pixel standard deviations to identify the signal base. Interpolation was then used to convert the signals to the target length. In comparison, [Yu et al. \(2024\)](#) (USST_Med) first used a YOLO network to detect image rotation. They used YOLO to detect lead names and then corrected the image based on the locations of these names. The image was subsequently converted to grayscale and divided into four segments, each of which was treated as a separate channel. They then used a ResUNet architecture for segmentation. To extract the signals, they grouped pixel values within each column. [Stenhede et al. \(2024\)](#) (Ahus AI Lab) first converted the image to grayscale before inversion. Similar to our approach, they used the Hough transform to detect grid lines and correct for rotations. They also used a U-Net model to segment the signals. To identify the leads, they assumed a fixed 3-by-4 layout.

Table 4: **Digitisation models tested in this work.** ‘SNR official’ is the SNR leaderboard score on the hidden set of the Challenge. ‘SNR CV’ is the average local score across cross-validation test sets. We applied different augmentation schemes to each of the three models during image segmentation. While the mask used for the M1 model was the original, the masks for models M2 and M3 were dense, comprising an interpolated set of pixels (see [Section 3.3.2](#)). Model M3 was first trained locally on 69k images and then further fine-tuned on 100 of the hidden Challenge train data.

Model	Train images	Augmentation	Mask	SNR official	SNR CV
M1	20k	Headers, calibration, wrinkles, rotation, cropping	Sparse	12.15	17.02
M2	45k	M1 + handwritten notes, QR-codes, var. colours	Dense	9.42	12.58
M3	69k + 100 fine	Same as M2	Dense	4.20	15.27

3.4 Results

3.4.1 Model performance

[Table 4](#) shows the results of different models for the digitisation task. Models M2 and M3, while in theory more advanced, were more prone to overfitting, requiring more training data to achieve a similar level of performance as M1. In the rotation step, we achieved a correct rotation prediction on over 99.7% of the images. Manually investigating a selection of the images where the rotation failed suggests that failure occurs when the colour schema and/or the augmentation are so extreme that determining the straight lines becomes very difficult. Therefore, vectorisation often did not capture the peaks of the signals, as illustrated in [Figure 10](#).

In addition to the PTB-XL scans used in the official ranking, models were evaluated on a variety of additional image types ([Reyna et al., 2024a](#)). Although we still achieved positive scores (0.51 to 4.93) on both colour and black-and-white scans of various paper qualities (deteriorated and clean), we scored negatively (-1.76 to -0.72) on mobile phone photos and computer screenshots.

3.4.2 Benchmark comparison

Compared to the four other papers achieving an SNR above 2, our approach achieved an SNR of more than twice that of the next best model ([Table 5](#)). Interestingly, each of

Table 5: **Benchmark comparison.** The table compares our model M1 with all other official submissions to the 2024 Challenge that achieved at least an SNR of 2.0. Reported is the ‘SNR official’, which is the leaderboard performance on the hidden set of the Challenge.

Model	Rotation method	Segmentation	Signal separation	SNR official
Our model (M1)	Hough transform	nnU-Net	Segmentation output	12.15
BAPORLab (Yoon et al., 2024)	na	U-Net	Post-processing rules	5.49
WAVIE (Verlyck et al., 2024)	Trained NN	nnU-Net	YOLO for bound. boxes	5.47
AhusAI (Stenhede et al., 2024)	YOLO	ResUNet	Post-processing rules	3.05
USST_Med (Yu et al., 2024)	Hough transform	U-Net	Post-processing rules	2.20

those models used an U-Net approach for segmentation, but with different pre-processing methods of the images and different vectorisation strategies for the extracted masks.

3.5 Discussion

3.5.1 Main findings

In this study, we combined several common image processing approaches, such as the Hough transform for line detection, U-Nets for segmentation and interpolation techniques for scaling. This method achieved the highest Challenge score, surpassing the next best-performing team by more than double. Our findings highlight that combining traditional image processing with deep learning models can effectively handle the inherent variability in paper ECG scans. Moreover, our pipeline demonstrates that robust digitisation methods, when applied systematically, can generalise well to diverse ECG layouts and recording conditions. These results reinforce the importance of integrating data-centric and learning-based approaches for digitising clinical waveforms.

3.5.2 Strengths of the study in relation to other studies

Our approach to ECG digitisation aligns with the multi-step approaches commonly reported in the literature (Baydoun et al., 2019; Fortune et al., 2021; Li et al., 2020b; Wu et al., 2022). Initially, we applied image pre-processing using the Hough transform (Baydoun et al., 2019; Wu et al., 2022). While many studies relied on pixel thresholds for segmentation (Baydoun et al., 2019; Fortune et al., 2021; Wu et al., 2022), our objective

was to improve generalisability by using a deep learning network (nnU-Net (Isensee et al., 2021)), similar to the approach in (Li et al., 2020b), followed by vectorisation. However, unlike this previous work (Li et al., 2020b), we applied our model to the entire image without prior pre-processing, with the exception of a rotation step. The key contributions of this approach are that it a) achieves much better results than other submissions to digitise ECG signals and b) can be easily applied to various data sources, whereas previous approaches are highly dependent on the specific properties the images used (e.g. assuming a particular grid colour). As long as the training data mimic the deployment settings, our approach can be easily adjusted.

When comparing our approach with other leading submissions (Table 5), all teams applied variants of the U-Net architecture for segmentation, alongside different pre- and post-processing steps. While all performed some form of rotation correction, some trained networks to detect rotation, whereas others used static methods such as the Hough transform. Additional pre-processing steps included greyscale correction (Stenhede et al., 2024; Yu et al., 2024), colour inversion (Stenhede et al., 2024) and segmentation of the image into multiple components (Yu et al., 2024). In contrast, we applied only rotation correction using the Hough transform. Although all teams used a U-Net-based approach for segmentation, the network architectures and training procedures varied. For example, Yoon et al. (2024) applied the architecture to simultaneously extract grid lines, text and waveforms using multiple channels, while this was applied in other studies only to segment the signals (Stenhede et al., 2024; Yu et al., 2024). The most substantial distinction in the digitisation process, however, was the extraction of individual signals. Whereas Verlyck et al. (2024) used a second network to identify bounding boxes and Stenhede et al. (2024); Yoon et al. (2024); Yu et al. (2024) used post-processing rules to separate individual signals, we extracted signals directly using the segmentation model. We propose that our network performed particularly well because it required minimal pre-processing, which makes it applicable to a broader range of images. We corrected image rotation using the Hough transform, which succeeded in 99.7% of cases. Furthermore, our vectorisation approach, which combined direct signal extraction from

segmentation with interpolation, proved especially efficient.

3.5.3 Limitations

Although the Challenge organisers used a hidden and diverse test set to encourage generalisable solutions, the public leaderboard only reflects performance on scans from the PTB-XL dataset. When evaluated on other sources, such as scanned and photographed images of varying quality (e.g. clean, stained, deteriorated), as well as screenshots of computer monitors, no team consistently achieved the highest performance across all datasets. In particular, performance on mobile phone photos of stained paper was notably poor, with no team obtaining a positive score. Moreover, our local cross-validation scores on augmented training data were substantially higher than the scores on the hidden test sets. This performance gap underscores the difficulty of achieving robust generalisation and points to the need for a training dataset that better represents the diversity of real-world input. Several assumptions in our approach also affected performance. Our training data contained signals arranged in a fixed order and at consistent positions. Although the segmentation model could, in principle, learn labels from the image content, it likely learned to rely on the consistent signal positioning instead. To mitigate this, we tested combining bounding box detection with optical character recognition (OCR) to increase spatial robustness, but this approach did not yield reliable results in local experiments. Further work is required to improve OCR integration and decouple signal recognition from spatial assumptions. Additionally, we assumed that each image contained only one complete signal, which does not hold for all cases. This assumption limited the ability of the model to generalise to images with multiple signals. A more robust solution would involve training the segmentation model on a broader set of examples that include multiple complete signals per image, enabling it to correctly handle a wider range of layouts.

3.5.4 Study implications

Developing methods to digitise printed or scanned images of waveforms can substantially support machine learning-based screening systems. Digitisation expands the range of available training data and enables the deployment of models in more diverse clinical and technical environments. However, achieving robust and generalisable performance remains challenging. To do so requires a data-centric approach and substantial resources, particularly to curate diverse and representative datasets that reflect real-world variability.

A modular approach that separates data pre-processing from prediction models offers several advantages. It allows researchers to focus on improving model performance without being constrained by specific input formats. Moreover, it facilitates the reuse and adaptation of trained models across different settings, as long as appropriate pre-processing steps are in place. This separation can promote reproducibility and broader applicability of models.

4 A generalisable approach for multimodal waveform integration

Extended chapter based on the paper (Krones et al., 2023)*:

Krones F, Walker B, Parsons G, Lyons T, Mahdi A. *Multimodal Deep Learning Approach to Predicting Neurological Recovery From Coma After Cardiac Arrest. Computing in Cardiology* (2023).

*Winner of the Best Preprint Award at the 2023 CinC conference for PhysioNet

 felixkrones/physionet_challenge_2023

Abstract. This chapter investigates how waveforms can be used in deep learning pipelines and showcases our team’s contributions to the 2023 George B. Moody PhysioNet Challenge. The aim was to integrate multi-channel EEG and ECG signals on the example use case of predicting neurological recovery from coma following cardiac arrest. Our modelling approach was multimodal, based on two-dimensional spectrogram representations derived from numerous waveform channels, alongside the integration of clinical data and features extracted directly from EEG recordings. The study shows the efficacy and limitations of using general purpose approaches in medical classification tasks. With regard to potential deployment, our analysis reveals that the performance of the model is strongly linked to the selection of a decision threshold, making an argument for clear uncertainty communication.

Contents

4.1	Introduction	78
4.2	Related work	80
4.3	Methodology	83
4.4	Results	87
4.5	Discussion	90

4.1 Introduction

Like the previous chapter, this work builds a foundational component for the overarching question of [Chapter 1](#). Where [Chapter 3](#) made paper ECGs machine-readable, this chapter asks how high-frequency waveforms can enter multimodal pipelines without expert feature engineering, addressing the lack of generalisable, scalable waveform integration highlighted in the review ([Chapter 2](#)).

Physiological waveform signals are an important feature in clinical decision-making across many different conditions ([Lopez Alcaraz and Strodthoff, 2024](#)), especially those that are cardiac related ([Reyna et al., 2023](#); [Strodthoff et al., 2024](#)). One of the most common cardiac diseases is cardiac arrest, with more than six million occurrences annually. General survival rates vary between 1% and 10% due to geographical disparities ([Reyna et al., 2023](#)). Following successful resuscitation, a substantial proportion of survivors are admitted to intensive care units (ICUs) in a comatose state, with severe brain injury emerging as the leading cause of death among this group. In the crucial days following cardiac arrest, medical professionals are often tasked with estimating the likelihood of the patient regaining consciousness. These predictions influence crucial treatment decisions, as a favourable outlook may support the continuation of care, whereas a poor prognosis often leads to the withdrawal of life-supporting care. However, there are documented cases of recovery despite early negative predictions, raising concerns about the risk of self-fulfilling prophecies ([Forgacs et al., 2020](#)).

Physiological waveform signals including Electroencephalograms (EEGs) and Electrocardiograms (ECGs) are crucial for assessing neurological and systemic function following cardiac arrest. Patterns within these signals, such as suppressed voltage, seizure-like activity, or heart rhythm disturbances, are known to correlate with neurological outcomes ([Hirsch et al., 2021](#); [Hofmeijer et al., 2015](#); [Reyna et al., 2023](#)). In clinical practice, these waveforms often inform decisions regarding prognosis, particularly in the critical days after return of spontaneous circulation (ROSC), when patients are frequently unresponsive.

While waveform signals are important in the prediction of many clinical conditions, interpreting such data requires expert review, which is labour-intensive and often unavailable outside specialised centres (Khazanova et al., 2021; Reyna et al., 2023). Automating waveform analysis has the potential to support more objective, timely and equitable prognostication across healthcare settings. Despite recent advances, predictive models based on physiological waveforms still face notable limitations. Most importantly, studies that report strong performance with multimodal data integration often rely on expert-curated, manually defined features tailored to specific datasets or tasks (Zabihi et al., 2023).

This has led to the following research question:

RQ: *Can spectrogram-based image models offer a generalisable alternative to expert-designed feature extraction for waveform-based tasks?*

To address this, as part of the PhysioNet Challenge 2023, we developed a multimodal deep learning framework that combines two-dimensional spectrogram representations derived from multi-channel waveform data with structured clinical variables and features extracted directly from the time series. Our approach is designed to operate on large-scale datasets and to model both the temporal evolution of signals and their interaction with clinical context. Using the newly released I-CARE dataset provided through the 2023 George B. Moody PhysioNet Challenge, which includes over 50,000 hours of recordings from more than 1,000 patients, we contribute a model to improve the reliability and applicability of automated prognosis for the example use case of comatose survivors of cardiac arrest. We compared our approach with the other top-performing submissions of the Challenge and showed that our highly generalisable method reaches 95% of the performance of a strongly expert-based approach.

The core contribution of this chapter is:

- **A spectrogram-based classification model:** We developed a multimodal framework that integrates spectrogram-derived features with clinical and time-series data, demonstrating its applicability to large-scale waveform tasks beyond expert feature engineering, using coma recovery prediction as a representative use case.

4.2 Related work

Physiological waveforms play a key role in clinical decision-making across a range of cardiac and non-cardiac conditions (Alcaraz et al., 2025), yet interpreting them demands specialised expertise, which is both time-consuming and frequently unavailable outside of specialist centres (Khazanova et al., 2021; Reyna et al., 2023). The automation of waveform analysis offers the promise of delivering more consistent, prompt and fair prognostic information across diverse healthcare environments. Although progress has been made, predictive models using physiological waveform data continue to encounter substantial challenges. A key issue is that many high-performing models, particularly those integrating multiple data sources, depend heavily on manually designed, expert-curated features tailored to specific tasks or datasets (Zabihi et al., 2023).

4.2.1 Automatic waveform analysis

Recent research has advanced automatic waveform analysis within clinical decision support systems. Studies have demonstrated that models leveraging physiological waveforms, such as ECG and phonocardiogram signals, can significantly outperform traditional diagnostic algorithms, particularly when combined with complementary data modalities. For example, deep learning models that integrate ECG and chest X-ray images have demonstrated higher accuracy in localising cardiac accessory pathways than traditional decision trees (Nishimori et al., 2021). Nishimori et al. (2021), for example, used an early fusion approach, in which an embedding vector derived from the X-rays was added as a 13th lead to the ECG data before training a one-dimensional CNN.

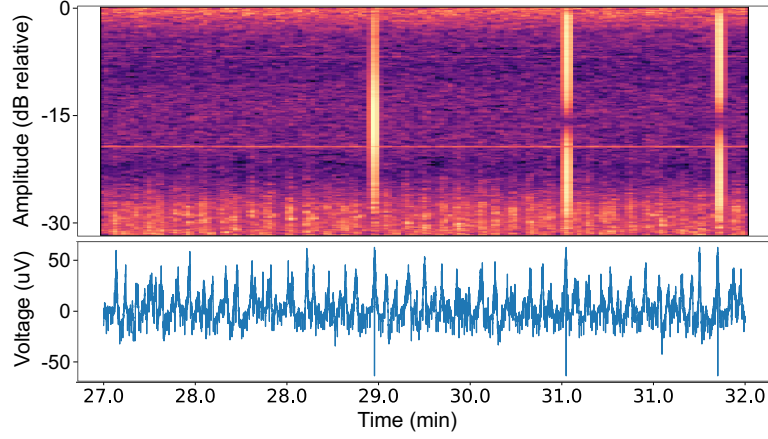


Figure 11: **Transformation example for patient 994.** *Bottom:* EEG recording of patient 994 at the 8th hour of channel F8 (right-frontal side) with good outcome. *Top:* Corresponding spectrogram (squared amplitude in decibel units relative to peak power).

Likewise, CNNs trained on log-Mel spectrograms of heart sound recordings, combined with patient demographics and signal-derived features, have improved heart murmur detection in paediatric populations (Walker et al., 2022). To this end, Walker et al. (2022) transformed the audio waveforms into spectrograms and used image models pre-trained on ImageNet (Deng et al., 2009). More recent efforts have shifted towards integrating raw waveform data with clinical variables to improve predictions of wider cardiac and non-cardiac outcomes. Multimodal models combining ECG waveforms with clinical data have achieved high predictive accuracy for conditions including cardiac arrest, respiratory failure and mortality, underscoring the added value of waveform data in capturing physiological changes (Lopez Alcaraz and Strodthoff, 2024). For example, Alcaraz et al. (2025) used waveforms to improve ICD-code and deterioration prediction. Additionally, image-based deep learning approaches have demonstrated the potential for digitising and classifying ECG images, addressing the gap between clinical practice and AI model design (Dias et al., 2024). However, despite these advances, challenges remain in generalisability, dataset representativeness and integration into routine care, indicating the need for further research and validation to fully harness the clinical potential of automatic waveform analysis (Strodthoff et al., 2024).

4.2.2 Outcome prognosis after cardiac arrest

A currently unresolved problem in waveform-based analysis is the prediction of neurological outcomes in comatose patients following cardiac arrest. In this context, physiological waveforms such as EEG and ECG recordings provide valuable information about cerebral function and the extent of brain injury (Reyna et al., 2023). Manual review of EEGs from 850 patients demonstrated that early EEG recordings can be reliable indicators of neurological outcome (Ruijter et al., 2019). While general advances in automated waveform analysis have established foundational techniques, this task remains particularly challenging and clinically important. Only a limited number of studies have focused specifically on outcome prediction after cardiac arrest.

CNN-based models have shown the potential to process EEG data more efficiently and consistently than human experts, with the possibility of enabling real-time decision support at the bedside. However, these models are typically trained on small datasets, which limits their robustness and generalisability (Khazanova et al., 2021; Reyna et al., 2023). For example, Tjepkema-Cloostermans et al. (2019) and Jonas et al. (2019) both applied VGG networks (Simonyan and Zisserman, 2014) to datasets with fewer than 700 patients. Multimodal approaches combining EEG with clinical variables may improve predictive performance, but a standard strategy for integrating such data is lacking. Zheng et al. (2021) proposed a model that combines EEG with demographic data using a CNN-LSTM architecture. CNNs were used to extract features from EEG segments, which were then passed across multiple time windows to LSTMs. A late fusion (cf. Chapter 2) method combined predictions from two LSTMs with outputs from a random forest trained on demographic variables. Other studies converted EEG signals into spectrograms and subsequently used CNN outputs and demographic data as inputs to a final random forest classifier. Despite promising results, many studies report only moderate sensitivity at clinically relevant specificity thresholds: sensitivity often remains below 70% when the false positive rate is constrained to less than 5% (Tjepkema-Cloostermans et al., 2019; Zheng et al., 2021). The most accurate models

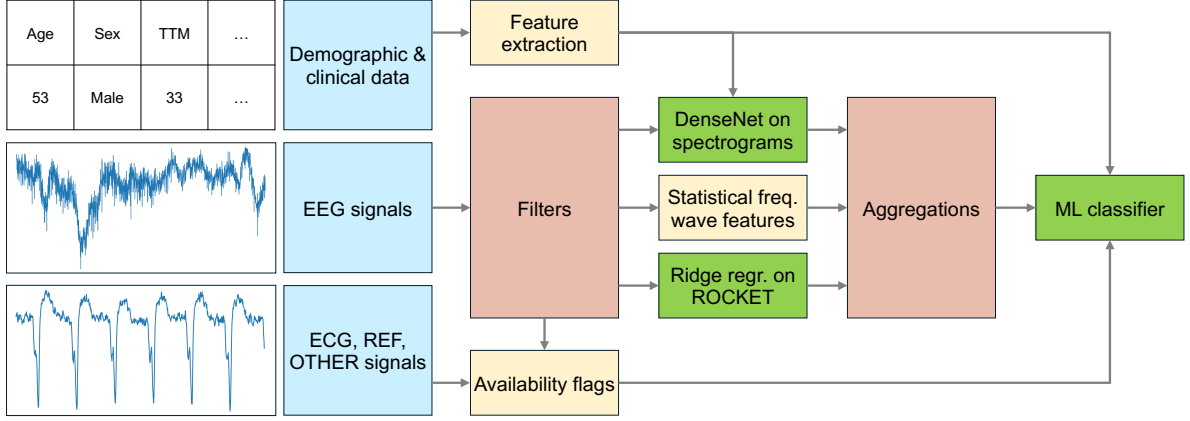


Figure 12: **A schematic diagram of our model architecture.** *Blue* blocks denote input data modalities, including demographic and clinical variables, multi-channel EEG signals and ECG or other physiological signals. *Salmon* blocks indicate signal filtering and aggregation operations applied across time and channels. *Yellow* blocks correspond to predefined feature representations, such as EEG summary statistics, frequency-band features and availability flags. *Green* blocks represent trainable models, including a DenseNet applied to EEG spectrograms, a ridge regression model on ROCKET features and the final random forest classifier.

typically rely on extensive expert-defined features; for instance, [Zabihi et al. \(2023\)](#) extracted over 300 such features. It remains unclear how generalisable multimodal deep learning frameworks must be designed to offer a viable alternative to expert-based feature engineering in waveform-based prediction tasks.

4.3 Methodology

4.3.1 Data

The dataset for this study is taken from the International Cardiac Arrest REsearch (I-CARE) consortium and originates from seven academic hospitals across the United States and Europe ([Amorim et al., 2023](#)). It comprises *clinical data*, including age, gender, return time of spontaneous circulation (ROSC), arrest location (out-of-hospital cardiac arrest OHCA), presence of shockable rhythm, use of targeted temperature management (TTM) and *clinical time-series data*, such as continuous EEGs, ECGs and other recordings (e.g. SpO2). The dataset consists of 1,020 patients (of which 607 were provided for training) who were admitted to an ICU in a comatose state following cardiac arrest. Neurological outcomes were assessed using the Cerebral Performance

Category (CPC) scale¹, derived from either phone interviews or chart reviews conducted 3-6 months post-ROSC. To ensure privacy, all clinical and EEG data were de-identified. EEG timestamps were organised based on the elapsed time since ROSC (Goldberger et al., 2024)

We filtered the EEG signals by applying a band-pass filter over the range 0.5–30 Hz and a notch-filter at 50 and 60 Hz to mitigate artifacts from electrical grids. We used artifact detection using a sliding window approach, only keeping the cleanest section of each recording, defined by a window size, stride size and epoch size. For each signal, we introduced a hyperparameter indicating whether to use either the entire signal or only the highest-quality window. The quality of each window was determined by examining each epoch for the presence of artifacts. The quality score of each window was the ratio of artifact-containing epochs to the total epochs. A hyperparameter was used to decide how many channels must contain artifacts for the epoch to be labelled as artifact-containing. Though our primary focus was on flat signal and amplitude threshold filters, abrupt slope changes, among other methods, could further refine artifact detection. Adjustments to signal lengths were made based on hyperparameters, which determined the minimum required signal length and the duration to be omitted at the start and end of each signal.

For part of our model (Section 4.3.2) we converted the EEG signals to spectrograms (Figure 11), using the Python package `librosa`, to use them together with other non-imaging modalities as inputs for the models (Walker et al., 2022).

4.3.2 Models and architecture

In our approach (Figure 12) we evaluated six models for binary prediction of patient outcome at 72 hours after ROSC. Model M1 used clinical features and EEG summary statistics before using a Random Forest classifier. In model M2, we added features

¹CPC = 1: good neurological function and independent for activities of daily living. CPC = 2: moderate neurological disability but independent for activities of daily living. CPC = 3: severe neurological disability. CPC = 4: unresponsive wakefulness syndrome [previously known as vegetative state]. CPC = 5: dead. (Reyna et al., 2023)

Table 6: **Summary of clinical data and patient outcomes.** Acronyms: ROSC = return of spontaneous circulation; OHCA = out-of-hospital cardiac arrests (True indicates OHCA); TTM = targeted temperature management (na indicates no TTM was applied).

Feature	Value	Missing
Age [years], mean (SD)	61 (16)	1
Sex ['Men'], N(%)	417 (69)	0
ROSC [minutes], mean (SD)	23 (19)	304
OHCA ['True'], N(%)	442 (78)	41
Shockable rhythm ['True'], N(%)	297 (52)	32
TTM, N at 33/36/na °C	448/61/98	0
Outcome ['Poor'], N(%)	382 (52)	0

extracted from EEG spectrograms using a DenseNet121 (Huang et al., 2017) into the classifier. In model M3, these DenseNet features were further aggregated over time and channels. Model M4 introduced an intermediate fusion step for clinical features into the last layer of the DenseNet architecture. In models M5 and M6, we introduce additional output from a ridge regression classifier. This classifier is trained on features extracted from the EEG signals using the Random Convolutional Kernel Transform (ROCKET) (Dempster et al., 2020). We obtain the regularisation strength of the classifier through leave-one-out cross-validation over 10 log-evenly spaced values ranging from 10^{-3} and 10^3 . ROCKET uses 10^4 kernels, with lengths randomly selected from the set $\{7, 9, 11\}$. Each kernel comprises 4 features and a maximum of 32 dilations. These settings align with the default parameters of the `RocketClassifier` from `sktime`. Model M5 omitted the intermediate fusion present in model M4, whereas model M6 included it.

4.3.3 Statistical analysis

The term significant denotes results with a p-value of less than 0.05. All statistical tests were two-sided. Continuous clinical variables were compared between independent outcome groups using Welch’s two-sample t-test, which does not assume homogeneity of variances. To assess performance differences between models evaluated on identical cross-validation folds, we used paired-samples t-tests on fold-level performance metrics, thereby accounting for the dependence induced by shared data splits.

Table 7: **Outcome frequencies by temperature management.** The table shows the outcome distribution obtained from the three different temperature management groups (33°C, 36°C and no management ‘na’).

Temp mgmt.	33°C (n=448)	36°C (n=61)	na (n=98)
Good	37.9%	39.3%	31.6 %
Poor	62.1%	60.1%	68.4%

4.3.4 Scoring

The official score of the Challenge emphasises the False Positive Rate (FPR) of incorrectly predicting a poor patient outcome. The scoring method selects the highest decision threshold that maintains the FPR below 5% and subsequently evaluates the True Positive Rate (TPR) for predicting poor outcomes. Mathematically the score is defined, given a threshold θ drawn from the predictions, as

$$\max_{\theta: \text{FPR}_{\theta} \leq 0.05} \text{TPR}_{\theta}. \quad (2)$$

Two types of scores were reported. The ‘Validation Score’ was the assessment provided by the Challenge organisers on the hidden set. The ‘CV Score’ and ‘CV AUC’ were computed as the mean and standard deviation, derived from the local five-fold cross-validation (CV).

4.3.5 Benchmark comparisons

To evaluate the performance of our spectrogram-based modelling approach, we compared it with the top-performing submissions from the Challenge. While our aim was to develop a generalisable method for incorporating waveforms into multimodal pipelines, the winning team adopted a highly tailored strategy called AIrhythm (Zabihi et al., 2023). They extracted more than 350 expert-defined features from the signals and used them in a dedicated classification model (Zabihi et al., 2023). Following several pre-processing steps, including band-pass filtering, Zabihi et al. (2023) extracted features from 3-minute waveform segments. These features were then aggregated across different hourly time points. Finally, they used an ensemble of tree-based methods, neural networks and

support vector machines for classification.

We also compared our method with the other submissions using a spectrogram-based approach (Bi et al., 2023; Li et al., 2023; Wu and Tan, 2023; Xia et al., 2023). The next best spectrogram-based approach was MelicientNet (Wu and Tan, 2023). Wu and Tan (2023) first converted the waveforms into spectrograms using the `librosa` package. They used two different EfficientNet CNNs (Tan and Le, 2019) to obtain predictions from the waveform data. These predictions were then combined with demographic and recording metadata and the complete feature set was passed to a random forest classifier. MetaHeart (Xia et al., 2023) used a 3D-ResNet architecture, having one spectrogram per hour and channel to predict the outcomes. MMCTNet (Bi et al., 2023) combined the output of a ResNet model on spectrograms with the output of a 1D-CNN on the EEG signals via an attention block before passing it to a classifier layer. Li et al. (2023) (UoMEEE) used a CNN on top of spectrograms and compared it with explicitly modelling the signals.

4.4 Results

4.4.1 Data analysis

Table 6 shows the feature availability and their overall key characteristics for the training data provided by the organisers of the Challenge. A subset of 607 (of total 1,020) patients was provided for training, which contained 32,712 hours (or partial hours) of data in 80,809 recording segments.

The patient outcomes are strongly associated with extreme CPC scores: for instance, 80% of patients with a good outcome had a score of 1, while 92% of patients with a poor outcome had a score of 5. While 69% of the patients were male, the distribution of outcomes was relatively consistent across gender, with 67% of female patients and 61% of male patients experienced poor outcomes. Patients with a poor outcome were significantly older on average (57.5 vs. 63.3). The average ROSC was 23.3 minutes. The maximum recorded ROSC time was 111 minutes, although this data was only available

Table 8: **Model performance.** The Val. Score is the performance on the hidden validation set of the Challenge. CV Score is the mean Challenge score from the five-fold cross validation on the training data with the standard deviation in parenthesis. CV AUC is the AUC on the same data. The ‘+’ should be read as ‘in addition’. The best performing model on the hidden validation set was M5. While M6 showed better performance locally, it was not evaluated on the hiddel validation set before the Challenge ended.

Model	Features	Val. Score	CV Score	CV AUC
M1	Clinical data, EEG summary stats, Signal flags	0.520	0.327 (0.137)	0.688 (0.046)
M2	+ DenseNet on spectrograms	0.540	0.484 (0.157)	0.824 (0.023)
M3	+ aggregation of features over time and channels	0.567	0.527 (0.090)	0.811 (0.049)
M4	+ intermediate fusion	0.328	0.537 (0.104)	0.818 (0.038)
M5	+ ROCKET features without intermediate fusion	0.627	0.447 (0.085)	0.836 (0.033)
M6	+ ROCKET features with intermediate fusion	not provided	0.567 (0.085)	0.854 (0.015)

for approximately half of the patients. The analysis did not identify a significant influence of temperature management on patient outcomes (Table 7).

4.4.2 Model performances

Table 8 provides the official scores on the hidden validation set, along with the mean and standard deviation for the score and AUC obtained from our local cross-validation, for the six models proposed in this work.

Figure 13 shows the relative feature importance of the classifier of our locally best performing model M6. The EEG signal information were the most important features while demographics and other clinical features exhibit lower importance. Figure 14 (left) shows the dependency of error rates on the selection of a particular threshold, with the ROC curve of our locally best performing approach (model M6) for which AUC=0.854 (Figure 14 (right)).

4.4.3 Benchmark comparison

The best-performing submission (Zabihi et al., 2023), based on more than 350 expert-designed features, achieved a Challenge score of 0.657, an AUC of 0.878 and an F1 score of 0.776 on the validation set. In comparison, our general approach (model M5) achieved a Challenge score of 0.627, an AUC of 0.828 and an F1 score of 0.763. These results demonstrate that our spectrogram-based pipeline achieves over 95% of

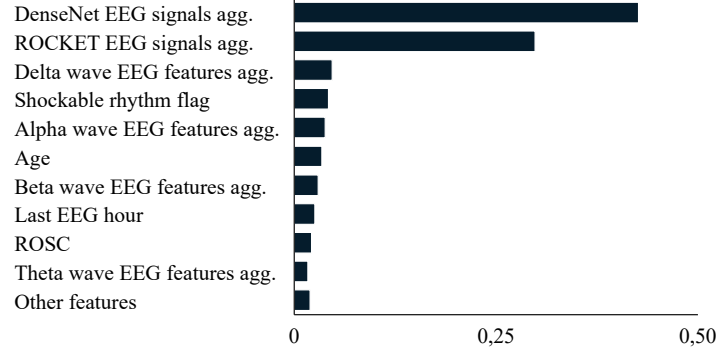


Figure 13: **Relative feature importance plot.** Shown is the feature importance of the locally best performing model M6. For signal-derived features, agg. indicates that features were aggregated across channels and time using the mean value. DenseNet features were aggregated through majority voting.

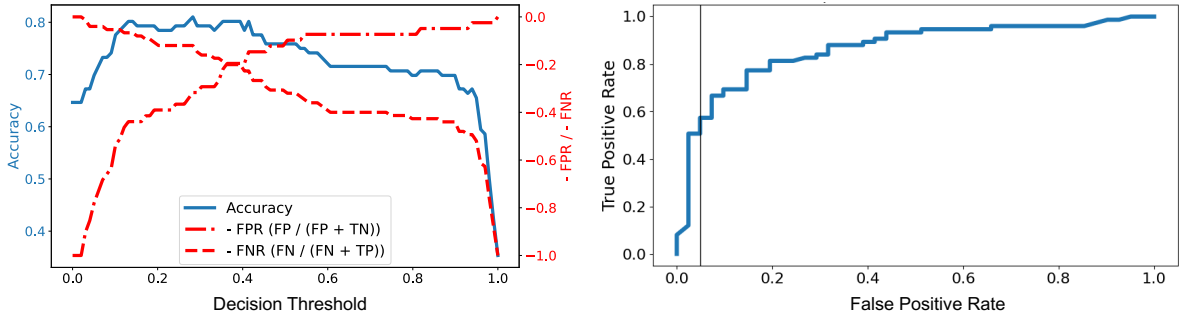


Figure 14: **Performance overview.** *Left:* Accuracy (blue), FPR (red dashed) and FNR (red dashed-dotted) of the ‘Poor’ outcome label for different decision thresholds of model M6. *Right:* ROC curve (blue solid) with the indicated 5% FPR threshold (black line).

Table 9: **Benchmark comparison.** The table compares our model M5 with the best performing model of the Challenge (Airhythm) and the other spectrogram-based approaches (MelicientNet, MetaHeart, MMCTNet, UoMEEE). Shown are the Challenge Score, AUC and F1-Score on the hidden validation set.

Model	Feature extraction	Classification	Score	AUC	F1
Our model (M5)	Spec. DenseNet, ROCKET	Early fusion, RF	0.627	0.828	0.763
Airhythm (Zabihi et al., 2023)	Expert-based	Early fusion, Ensemble	0.657	0.878	0.776
MelicientNet (Wu and Tan, 2023)	Spec. EfficientNet	Late fusion, RF	0.597	0.871	0.737
MetaHeart (Xia et al., 2023)	Spec. 3D-ResNet	ResNet	0.582	0.829	0.703
MMCTNet (Bi et al., 2023)	Spec. ResNet + 1D-CNN	Attention, FC layer	0.567	0.828	0.749
UoMEEE (Li et al., 2023)	Spec. CNN	CNN	0.388	0.629	0.385

the performance of the highly specialised expert-feature model, while offering greater flexibility for broader applications (Table 9). The next best spectrogram approach (Wu and Tan, 2023) achieved a Challenge score on the validation data of 0.597, an AUC of 0.871 and an F1 score of 0.737.

4.5 Discussion

4.5.1 Main findings and limitations

In this study, we developed and tested six models for forecasting post-cardiac arrest coma recovery using a multimodal approach. Model M1 served as our benchmark. Incorporating CNN-extracted features from spectrograms in M2 led to the most significant AUC improvement in local CV. To improve robustness, we aggregated features from different hours and channels in model M3. This aggregation effectively reduced the number of features used in the final classifier and improved the score, which was highly sensitive to minor variations on the left side of the AUC curve (Figure 14). Although model M4’s intermediate fusion improved local score, its validation performance declined, indicating overfitting. Adding features extracted by ROCKET to the final classifier substantially improved performance on the validation set (0.627), but had limited impact on the local CV set. Furthermore, performance deteriorated on the hidden test set (0.53), leading us to conclude that enhancing the robustness of our methodology remains an open research question. Additionally, given that ROCKET features did not markedly improve local results, it is still unclear how much additional information these features provide compared to those already extracted. Our results show that although multimodal approaches (M2-M6) achieved strong performance on the local held-out set (CV AUC > 0.81), generalisability concerns emerged, which need further considerations.

It is worth stressing that the strongest absolute performance was still reached by the expert-based approach (Zabihi et al., 2023), which confirms that domain expertise remains important for waveform-based prognosis. Expert-engineered features also carry a clear advantage in interpretability, as each feature has a physiological meaning and encodes prior expectations about how the signal reflects the underlying condition, which makes their behaviour easier to explain to clinicians and, in some settings, lets them match or beat end-to-end models. The contribution here is therefore not that expertise can be replaced, but that the central challenge is one of generalisability. Expert-curated features are tied to specific datasets, tasks and pre-processing choices, whereas

a spectrogram-based approach trades a small drop in performance for a method that transfers more easily across settings.

A related question is which modalities drive the prediction and whether a multimodal model is needed for this task at all. The feature importance analysis (Figure 13) showed that EEG-derived features dominated, while demographic and other clinical variables contributed less. This pattern is expected on clinical grounds, as neurological recovery after cardiac arrest is determined mainly by the degree of hypoxic brain injury, which the EEG reflects most directly, whereas variables such as age or ROSC are only indirect proxies. Early EEG is accordingly a known reliable indicator of outcome (Hofmeijer et al., 2015; Ruijter et al., 2019). The clinical and demographic data added only modest value over the EEG-derived features, and the ECG signals could not be fully exploited because of substantial missing data. For coma prognosis, the predictive signal thus came largely from a single modality. We still consider the multimodal framing worthwhile. The additional modalities improve robustness when EEG quality is poor, and the pipeline is designed to transfer to tasks where other modalities carry more weight.

4.5.2 Strengths of the study in relation to other studies

This study demonstrates that generalisable multimodal models can approximate the performance of highly specialised pipelines in the context of waveform-based clinical outcome predictions. Compared to the winning submission of the PhysioNet Challenge (Zabihi et al., 2023), which extracted more than 350 expert-curated features using extensive pre-processing and tailored classifiers, our spectrogram-based approach achieved over 95% of their performance without relying on expert domain knowledge. This suggests that our model may generalise more easily across settings with differing pre-processing protocols or subject-specific expertise. Furthermore, our approach also makes it easier to use and maintain without expert help.

Several other submissions adopted spectrogram-based approaches (Bi et al., 2023; Li et al., 2023; Wu and Tan, 2023; Xia et al., 2023). In contrast, a few teams, such as Khambampati et al. (2023), also extracted spectral features but used them as input to

tree-based models, for example, XGBoost. Among the better-performing submissions (ranked within the top 10), most used CNNs on pre-processed signals (Kim et al., 2023; Zhou et al., 2023a), or combined multiple feature extraction strategies, such as CNNs, tree-based models and support vector machines, via fusion or ensembling techniques (Patidar and Sawant, 2023; Uslenghi et al., 2023; Yang and Tetzlaff, 2023). Recurrent architectures were also used, as in the LSTM-based approach by Hempel et al. (2023). Overall, no single method consistently outperformed others and no clearly superior strategy emerged.

Unlike many prior studies in coma prognosis which focused solely on EEG data and smaller cohorts (Jonas et al., 2019; Tjepkema-Cloostermans et al., 2019), our work is based on one of the largest multimodal datasets to date (I-CARE), with over 1,000 patients and 50,000 hours of waveform data. This strengthens the reliability of our findings and facilitates model training on a wider range of signal durations, noise levels and channel configurations, potentially improving applicability across clinical environments.

Unlike Zheng et al. (2021), who used a late-fusion architecture combining CNN-LSTM pipelines with a random forest on features requiring expert input, we used spectrogram-based representations as a unified input format. Furthermore, we showed that dense convolutional features derived from spectrograms are among the most predictive inputs (Figure 13), even in the presence of clinical variables, indicating that these models effectively extract clinically relevant patterns from EEG signals.

In addition, our approach supports integration of complementary signal encodings, such as features extracted using ROCKET, further demonstrating the flexibility of the proposed pipeline.

Taken together, this study highlights that spectrogram-based models can offer a scalable alternative to hand-crafted pipelines, bridging the gap between task-specific expert models and general-purpose AI architectures suited for multimodal clinical data.

4.5.3 Future research

Several potential avenues for future research are available. For instance, one can consider refining the modelling of time dependencies (Zheng et al., 2021) and exploring self-supervised pre-training to boost network performance. The incorporation of other time-series data, such as ECGs (not fully exploited in this study due to substantial missing data) could also lead to valuable insights, particularly when integrated with EEG signals (Zabihi et al., 2023).

4.5.4 Study implications

This study demonstrates that spectrogram-based deep learning models provide a scalable and generalisable approach to incorporating waveform data into multimodal clinical decision-making. By transforming raw waveform signals into two-dimensional representations, deep learning architectures can autonomously extract relevant features, reducing dependence on manual feature engineering. This integration allows the combination of diverse data types, including physiological signals and structured clinical variables, to support timely and consistent clinical decisions.

5 Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders

This chapter is based on the paper [\(Krones et al., 2026\)](#):

Krones F, Papastylianou T, Qian G, Parsons G, Shukla B, Papież BW, Mahdi A. *Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders*. **Information Fusion** (2026) 136:104573.

Abstract. Although ICUs generate rich multimodal data, most studies examine only narrow modality combinations, limiting progress toward clinically useful decision-support systems. We propose a systematic framework for evaluating multimodal fusion strategies in the ICU, enabling structured comparisons of early, intermediate and late fusion across combinations of imaging, tabular, text and time-series data. To ensure generalisability, we apply the framework to six cardiovascular conditions covering a broad diagnostic spectrum. Rather than asking whether multimodal fusion helps, we ask under which conditions it helps. Multimodal models frequently outperform unimodal baselines, with gains of up to five percentage points in AUC, particularly when modalities provide complementary information or prediction uncertainty is high. Across the architectures we explored, late fusion is the most robust and best-performing strategy. Using uncertainty-based ranking, we show that multimodal integration is especially effective for correcting high-uncertainty cases, substantially reducing both false-negative and false-positive rates relative to single-modality models. The most informative data are typically available within the first 24 hours of ICU admission. These findings offer guidance for prioritising modalities and deploying multimodal models in real-world clinical settings.

Contents

5.1	Introduction	95
5.2	Methodology	100
5.3	Results	115
5.4	Discussion	125

5.1 Introduction

This chapter turns to the core of the overarching question set out in [Chapter 1](#), namely under which conditions multimodal fusion improves the prediction of cardiovascular conditions in the ICU. It addresses the central gap identified in the review ([Chapter 2](#)), the absence of a systematic comparison of fusion strategies across the modalities relevant to intensive care, and builds on the waveform representations developed in [Chapter 3](#) and [Chapter 4](#).

Background. Cardiovascular diseases remain the leading cause of death worldwide ([World Health Organisation, 2023](#)), responsible for approximately 17.9 million deaths each year, accounting for 32% of global mortality ([World Health Organisation, 2021](#)). Importantly, up to 80% of these deaths could be prevented through effective risk management ([Raleigh et al., 2022](#)). Artificial intelligence (AI) has demonstrated considerable potential to improve healthcare by increasing diagnostic accuracy and efficiency, particularly in critical care settings ([Moor et al., 2023](#); [Rajpurkar et al., 2022](#)). This is especially important as cardiovascular diseases account for a substantial share of ICU admissions ([Anandaciva, 2020](#); [Cook et al., 2022](#)).

ICU decision support. Decision-making in the ICU involves integrating multimodal information from laboratory results, imaging and video data, continuous monitoring signals and clinical documentation ([Amal et al., 2022](#)). Nevertheless, most existing AI models are limited to a single data modality ([Huang et al., 2020b](#)), which does not reflect how clinicians synthesise information from multiple sources. To be incorporated in real-time hospital environments, clinical decision-support tools must be able to process multimodal data types ([Gonzalez et al., 2024](#)). The limited availability of AI models with such capabilities limits their deployment to clinical settings. Hence, despite increasing interest in applying AI to intensive care and the increased interest in foundation models for medicine ([Moor et al., 2023](#)), many existing approaches remain poorly aligned with clinical workflows ([Pinsky et al., 2024](#)) and have not yet been widely adopted. Recent work on transformer-based foundation models trained on longitudinal electronic health

records has shown that zero-shot patient trajectory models can accurately predict ICU mortality, length of stay and readmission risk directly from patient health timelines, highlighting the potential of generalist decision-support systems in critical care (Renc et al., 2024). In practice, ICUs already rely on various scoring systems that use pre-selected variables to generate simple risk metrics. Examples include **S**equential **O**rgan **F**ailure **A**ssessment score (SOFA) (Vincent et al., 1996) for organ failure and **A**cute **P**hysiology **A**nd **C**hronic **H**ealth **E**valuation (APACHE) IV (Zimmerman et al., 2006) for length of stay and mortality predictions. However, these scores are often applied as broad prognostic tools rather than for condition-specific decision support and focus on a predetermined selection of variables that have been flagged clinically important, rather than evaluating the whole spectrum of available clinical information and allowing the AI model to dynamically determine feature importance.

Data availability. Especially the application of multimodal prediction models to cardiovascular conditions remains under-explored (Krones et al., 2025), largely due to the lack of suitable public data. The first publicly available datasets focusing on cardio related conditions, such as NIH Chest X-ray (Summers, 2019), CheXpert (Irvin et al., 2019) and PadChest (Bustos et al., 2020), were typically limited to X-rays and basic demographic information. Thus, early multimodal studies in this area were smaller and limited in the number of modalities they could incorporate (Bagheri et al., 2020; Li et al., 2021). Important progress occurred in 2021 with the release of MIMIC-IV (Johnson et al., 2023b), a large multimodal dataset comprising over 200,000 ED patients and 65,000 ICU patients. For the first time, this enabled the linkage of different kinds of tabular data, which in the context of this paper we will be referring to as LVD (for **L**ab tests, **V**ital signs and **D**emographics), with corresponding images (Hayat et al., 2022; Soenksen et al., 2022). In an updated version of MIMIC-IV in 2023, more data formats, such as ECGs and patients’ clinical notes, were made available, allowing further linkage.

Fusion comparison. Given the recent availability of this multimodal dataset and the advent of foundation models for healthcare (Moor et al., 2023), the question of how to best fuse different modalities becomes critical, as effective integration determines

whether models can fully exploit complementary information and generalise across clinical contexts. Data can be linked at different stages. The three main techniques correspond to when data fusion takes place and are typically referred to as *early*, *intermediate* and *late* fusion (Baltrušaitis et al., 2018; Huang et al., 2020b). While *early* fusion uses the outputs of frozen feature extraction models as input to a final prediction model, *intermediate* fusion backpropagates the loss function through the entire network. In contrast, *late* fusion aggregates the predictions of individual models. While each of these techniques has been applied separately to classify niche applications in healthcare, few studies have systematically compared the performance of all three fusion approaches on the same dataset in a clinical context (Huang et al., 2020c; Imrie et al., 2025). Consequently, there appears to be no consensus on an optimal fusion approach for multimodal data (if such an approach does exist) and, in particular, an approach relevant to data collected for clinicians in the ICU. While comparisons of different approaches have been undertaken by a few groups (Table 10), these studies typically involve either a limited number of modalities or focus on modalities that are less relevant in an ICU setting.

Complementary to the question of where modalities are fused, a separate line of work studies how to fuse them through information-theoretic objectives, which are loss functions that quantify how much information is shared between modalities through measures such as mutual information, total correlation or maximal correlation. For example, Huang et al. (2020d) learn feature representations across multiple variables by maximising their total correlation, while Mei et al. (2023) build on this idea for multimodal brain tumour segmentation, where features from four MRI sequences are combined at an intermediate stage through a partial common information mask and a self-attention module. A recent survey by Vavekanand et al. (2026) provides a broader synthesis of 52 medical multimodal fusion studies, reporting intermediate fusion as the most commonly adopted strategy (60% of studies), an average AUC gain of 5 to 12% over unimodal baselines and that only 12% of studies perform true external validation.

Gaps. The lack of a systematic evaluation of fusion techniques on ICU-relevant modalities

Table 10: **Comparison of exemplary fusion approaches.** Overview of papers on multimodal fusion in healthcare, showing that no consensus exists on a best-performing approach across all clinical applications of AI. The number of subjects is the number of patients reported in each study (1,837 studies for [Huang et al. \(2020c\)](#), 2,298 images of 1,641 lesions for [Imrie et al. \(2025\)](#) and 10,185 examinations for [Holste et al. \(2021\)](#)), except for [Hayat et al. \(2022\)](#), where it is the number of MIMIC-IV ICU stays in the phenotyping training split, of which 7,756 were paired with a chest X-ray.

Study	Application	Best approach	Modalities	No. of subjects
Huang et al. (2020c)	Pulmonary embolism classification	Late	Clinical patient data, CT	1,794
Imrie et al. (2025)	Lesion classification	Late/Ensemble	Images, tabular data	1,373
Guo et al. (2019)	Tumor regions segmentation	Early	PET, CT, MRI	50
Holste et al. (2021)	Breast cancer classification	Early/Intermediate	Clinical patient data, MRI	5,248
Hayat et al. (2022)	Mortality & phenotyping	Intermediate	Clinical data, X-rays	42,628

leaves a considerable gap in our understanding of the usability of such approaches to support ICU clinicians in decision-making. The main gaps in the literature that need to be addressed include a systematic comparison of the performance of different data fusion techniques on the many data modalities relevant to patients in the ICU and, subsequently, investigating the contribution of individual data modalities to predictive performance across a range of ICU-related conditions. These are all necessary prerequisites to build robust, clinically reliable foundation models for decision-support in the ICU. In addition to integrating information from multiple modalities, another major obstacle to clinical implementation in the ICU lies in accounting for the staggered availability of clinical data ([Sivaraman et al., 2023](#)). Despite this, most research on multimodal fusion assumes that all input features are available simultaneously ([Huang et al., 2020c](#); [Imrie et al., 2025](#)). Among the studies that do account for the temporal dynamics of data availability, the focus is often restricted to a single modality ([Deasy et al., 2020](#); [Kim et al., 2019](#); [Thiele et al., 2025](#)). Few approaches, however, address dynamic data acquisition across multiple modalities; one such study aimed to predict mortality and other outcomes after ICU admission ([Placido et al., 2023](#)).

To this end, the main research questions posed in this chapter are as follows:

RQ 1: *Does any fusion method consistently achieve superior performance in multimodal prediction of cardiopulmonary conditions in the ICU?*

RQ 2: What is the contribution of each individual data modality to predictive performance for cardiopulmonary conditions in the ICU?

RQ 3: At what point during a patient's stay in the ICU has data been sufficiently accumulated to make reliable predictions?

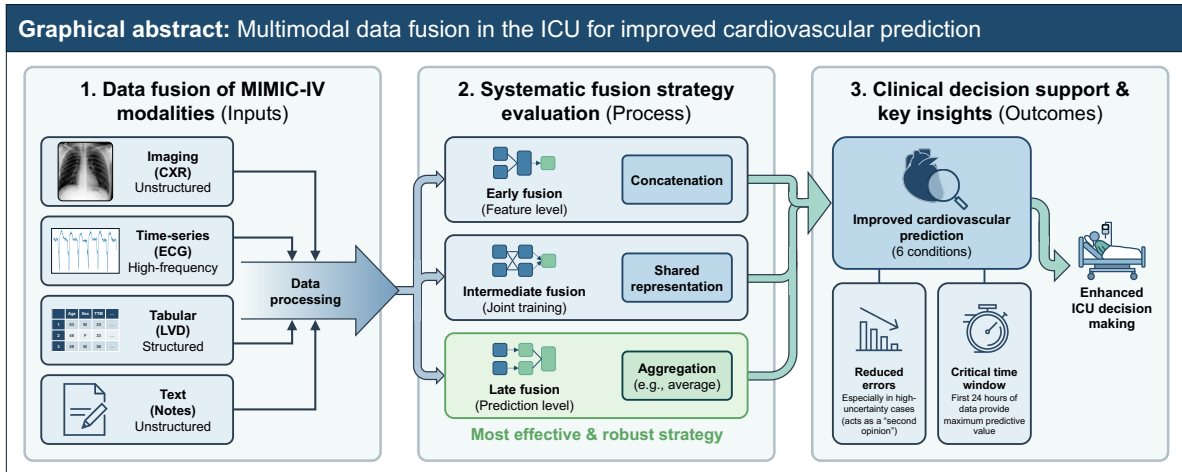


Figure 15: **Graphical abstract.** Multimodal data fusion framework for cardiovascular prediction in the ICU. 1. Input modalities derived from the MIMIC-IV database, comprising chest X-rays (CXR), electrocardiograms (ECG), tabular data (LVD) and clinical notes. 2. Systematic evaluation of three fusion strategies: early (feature level), intermediate (joint training) and late fusion (prediction level). 3. Key study outcomes identify late fusion as the most effective strategy, yielding improved prediction across six cardiovascular conditions, reduced diagnostic errors and highlighting the first 24 hours as the critical time window for data acquisition.

The core contributions of this chapter are:

- **Data processing and exploration:** We created a comprehensive and efficient pipeline for MIMIC-IV processing, integrating over 100 tables into a harmonized, event-based structure with a hybrid storage approach, alongside interactive dashboards and LLM-assisted documentation to enable intuitive exploration for interdisciplinary collaboration.
- **Systematic comparison:** We conducted the first large systematic comparison of early, intermediate and late fusion strategies on ICU-relevant clinical data modalities (imaging, time-series, tabular and text data). In contrast to prior work that has typically evaluated a single fusion strategy on a single condition, our framework allows controlled comparisons across strategies, modalities and conditions on the same dataset. Across the architectures we explored, late fusion was the most robust strategy, particularly when data availability was limited.
- **Contribution of modalities to model accuracy:** We introduced an ablation-based analysis that offers new insights into decision-making processes in the ICU. Using uncertainty-based ranking we show that multimodal data improve model performance especially for the most uncertain predictions and when additional modalities offer non-redundant information. We also find that additional modalities can degrade performance when they introduce mostly noise. Moreover, we show that in a large percentage of cases, the multimodal model results in a correct prediction where at least one of the single modalities would have produced a false positive result.
- **Model performance over time:** We evaluated how the overall predictive performance of the model and the relative importance of each feature changes over the course of a patient’s ICU stay. We show that especially the data collected during the first 24 hours after ICU admission improve predictive accuracy substantially.

5.2 Methodology

This section describes in detail the analytical methods used, following the adjusted TRIPOD+AI checklist ([Collins et al., 2024](#)).

5.2.1 Data

The predictive modelling of cardiovascular conditions in the ICU requires a structured and comprehensive dataset that integrates multiple data modalities ([Amal et al., 2022](#);

Duvieusart et al., 2022). However, as highlighted by Kansal et al. (2025), only very limited data sources provide the option to link image data, biological signal data, tabular data and free-text notes, i.e., MC-MED for ED patients (Kansal et al., 2025), ADNI for dementia detection (Weiner et al., 2017) and MIMIC-IV for ICU patients (Johnson et al., 2024).

This is why this study used data from the MIMIC-IV (version 3.0) database (Goldberger et al., 2024; Johnson et al., 2024, 2023b). MIMIC is a collection of freely available, de-identified medical data organised in a relational database. It contains de-identified health records of patients treated at the Beth Israel Deaconess Medical Center in Boston, USA, between 2008 and 2022. Of the 364,627 patients included in MIMIC-IV, approximately 65,000 were recorded as having been treated in the ICU. Overall, MIMIC-IV contains records of 546,028 hospitalisations for these 364,627 patients and 94,458 ICU stays for the 65,000 ICU patients.

5.2.2 Cohort

Eligible patients included those who were 18 years or older and first admitted to the emergency department, before progressing to the ICU. Only the first stay in the ICU was included in the analysis, contingent on a duration of stay of at least 24 hours. Any subsequent ICU stays were excluded. Patients who were not admitted to the ICU, not recorded as being discharged, or died during their first ICU stay were also excluded. After applying these criteria, the final cohort consisted of 28,262 patients (Table 27).

5.2.3 Data preparation

Despite the richness of MIMIC-IV, its raw format presents challenges for direct use in predictive modelling. Data are distributed across multiple tables, often requiring complex preprocessing steps to ensure alignment across modalities. Furthermore, missing values, temporal misalignment and variations in data quality introduce additional complexities that must be addressed to extract meaningful clinical insights. Various attempts have been made to combine the diverse types of data in MIMIC-IV and while these efforts

have contributed valuable milestones, they have proven insufficient in addressing the full range of multimodal challenges, highlighting the need for a new processing pipeline. Key contributions published through the official <https://physionet.org/> platform include HAIM-MIMIC-MM (Soenksen et al., 2022), a software used to create a multimodal dataset by integrating MIMIC-IV v1.0 and MIMIC-CXR-JPG v2.0.0, filtering patients to those with at least one chest X-ray. However, it suffers from several limitations: it applies fixed patient inclusion criteria, reducing flexibility in cohort definition and its applicability to research questions requiring alternative eligibility; relies on outdated requirements, such as Python 3.6 and `pandas` v1, specifically for a 64-bit macOS architecture, making it difficult to extend or reproduce the pipeline on contemporary systems; and excludes ECG waveform data, limiting its use in cardiovascular research (Soenksen et al., 2022). Similarly, the MIMIC-IV-Ext-MDS-ED dataset (Lopez Alcaraz and Strodthoff, 2024), focused on the first 1.5 hours after patient arrival, offers valuable data such as demographics, vital signs, lab results and ECGs, but is constrained by a limited data collection window and the absence of imaging data, such as X-rays. Additionally, it is restricted to emergency department data, making it less applicable for ICU-specific research. The Symile-MIMIC dataset (Saporta et al., 2025), while incorporating X-rays, ECGs and blood lab tests, suffers from the lack of ICU-specific data, as its inclusion criteria are based on time after hospital admission and not necessarily on critical care and only includes admissions containing all three modalities. The mentioned limitations have motivated the development of a more flexible and comprehensive pipeline to address the evolving demands of multimodal healthcare research.

While basic data cleaning steps have not been performed by the authors of MIMIC-IV to keep the data as close as possible to real-world datasets, patient identifiers were removed as part of de-identification and dates and times were randomly shifted. Moreover, all patients older than 89 were grouped together as being 91. We further preprocessed the data through a modular pipeline designed to integrate multimodal data sources from the MIMIC-IV database. Tabular data, both raw and processed, were merged into a

single event-based table. The resultant dataset included laboratory tests, vital signs and demographic information (LVD) (among other information) for each patient, which were then linked to their X-ray images, ECG waveforms and patient notes. Temperatures were all converted to the same unit. Body temperatures below 50 degrees Fahrenheit were removed. Diastolic blood pressures above 400 mmHg were removed.

Imaging data (chest X-rays) were linked to some of the MIMIC-IV patients using MIMIC-CXR (Johnson et al., 2019b). MIMIC-CXR-JPG v.2.1.0 (Johnson et al., 2019a) was used for the experiments in here. This includes 377,110 images from 227,835 studies of 65,379 of the MIMIC patients. ECGs were drawn from the MIMIC-IV-ECG dataset v1.0 (Gow et al., 2023). This dataset contains 800,035 ECGs (all 500Hz, 12-lead and 10 seconds in duration) from 161,352 of the MIMIC patients. The data are grouped for this work into tabular data which we call LVD data (admission, administration, treatments, diagnosis codes, static patient information, low-frequency measures), text data (discharge notes, measurement and treatment comments), high-frequency time-series data/waveforms (ECGs) and images (chest X-rays). Vital signs, such as heart rate, blood pressure, etc. are grouped into low-frequency measures and counted towards tabular data, as they are only stored aggregated.

For multimodal fusion, only data containing all required modalities were used. However, the pipeline was created so that it could still operate on samples with missing modalities. To do this, an identifier vector was added to identify missing modalities, which were replaced with zeros. For missing tabular features this was done by adding a binary mask feature for all clinical features and for modalities this was done by adding a binary value for each modality before fusion. As the fusion experiments were restricted to samples with all modalities available, modality-level zero-imputation was inactive in this study and the mask was retained only for forward compatibility with future analyses on cohorts with missing modalities. For tabular LVD data, a similar approach was used. While XGBoost supports missing values by default, for the use of neural networks, missing values were replaced with column means and a mask was added to highlight any missing values. This step is particularly relevant in clinical contexts, as the absence

of measurements may itself convey information: for instance, certain data are recorded only when the clinician suspects a particular condition.

5.2.4 Predictors

For LVD data, features were retained if they were present in at least one-third of all patients or had been included in prior studies (Duvieusart et al., 2022; Grant et al., 2021). For low-frequency time-series data such as temperature measurements, various aggregation methods were used, as described in Table 29 and Table 30. Unless otherwise stated, we used the final ECGs and X-ray images taken during the considered patient’s stay, as our objective was to benchmark the predictive performance achievable when all available in-stay data are used, independent of timing.

To avoid target leakage, clinical notes were first split into sections using OpenAI’s gpt-4.1-mini-2025-04-14 model via the OpenAI API, chosen for its speed and cost-effectiveness at the time of the experiments. The split used a structured-output prompt with a Pydantic schema (D.3), so that each fragment was assigned to one of 13 predefined sections rather than rewritten. In a second stage, all text related to the *brief hospital course* or *discharge* was removed, ensuring that no post-admission outcome information was available to the prediction models. Due to resource constraints, only the 5280 notes from the fusion dataset were cleaned. A subset of around 20 cleaned notes was additionally inspected manually against their raw counterparts; in this sample we did not find evidence of leaked outcome information. As we could not completely rule out the potential of target leakage from the remaining information, we kept reporting the results including notes separately, so that none of our headline claims depend on this modality.

The following modality-specific preprocessing steps were applied. *Chest X-rays* were resized to 224×224 pixels with random cropping, rotation and horizontal flipping applied as data augmentation during training. Pixel values were normalised using ImageNet channel-wise means and standard deviations. *ECG recordings* (all 500 Hz, 12-lead, 10seconds) were converted into mel-spectrograms using the `librosa` library,

Table 11: **Distribution of target labels across modalities.** The rows correspond to the six conditions considered in this study; each column to the right shows how positive and negative cases were distributed across all observations within that modality and for the respective condition. 'Fusion' denotes all cases for which all four modalities to the right were present.

Label	Fusion	LVD	X-rays	ECGs	Text
All	5,280	28,262	110,852	298,270	202,358
(1) Hypertensive diseases - 1	3,413 (65%)	17,455 (62%)	69,322 (63%)	199,148 (67%)	106,077 (52%)
(1) Hypertensive diseases - 0	1,867 (35%)	10,797 (38%)	41,530 (37%)	99,122 (33%)	96,281 (48%)
(2) Ischaemic heart diseases - 1	915 (17%)	3,271 (12%)	12,006 (11%)	52,972 (18%)	9,102 (4%)
(2) Ischaemic heart diseases - 0	4,365 (83%)	24,981 (88%)	98,846 (89%)	245,298 (82%)	193,256 (96%)
(3) Chronic i. heart disease - 1	1,474 (28%)	6,239 (22%)	28,720 (26%)	112,230 (38%)	37,407 (18%)
(3) Chronic i. heart disease - 0	3,806 (72%)	22,013 (78%)	82,132 (74%)	186,040 (62%)	164,951 (82%)
(4) Cardiomyopathies diseases - 1	251 (5%)	1,048 (4%)	6,000 (5%)	18,108 (6%)	4,937 (2%)
(4) Cardiomyopathies diseases - 0	5,029 (95%)	27,204 (96%)	104,852 (95%)	280,162 (94%)	197,421 (98%)
(5) Dysrhythmias diseases - 1	2,007 (38%)	8,856 (31%)	39,745 (36%)	120,537 (40%)	36,709 (18%)
(5) Dysrhythmias diseases - 0	3,273 (62%)	19,396 (69%)	71,107 (64%)	177,733 (60%)	165,649 (82%)
(6) Heart failure - 1	1,532 (29%)	6,279 (22%)	32,429 (29%)	95,753 (32%)	31,515 (16%)
(6) Heart failure - 0	3,748 (71%)	21,973 (78%)	78,423 (71%)	202,517 (68%)	170,843 (84%)

and the resulting spectrograms were treated as images and resized to 224×224 pixels, following the same augmentation pipeline as the X-ray data. *Tabular LVD data* were aggregated per patient using the methods described in Table 29 and Table 30; for neural network models, missing values were imputed with column means and a binary mask was appended to indicate missingness, while XGBoost handled missing values natively. *Clinical text notes* were tokenised using the ClinicalBERT tokeniser with a maximum sequence length of 512 tokens for the fine-tuned model, or passed as plain text to the GPT-4.1-mini API for zero-shot classification.

5.2.5 Outcome

The objective of this study was to predict cardiovascular conditions, identified using ICD-10 diagnosis codes assigned at ICU discharge. Each was modelled as a binary classification task. The selection and grouping of these diagnoses were informed by clinical relevance, label frequency and shared risk factors and refined after consultation with an ICU clinician, as detailed in Table 26. The codes were grouped using the latest WHO classification². To maintain analytical consistency, ICD-9 codes were mapped to ICD-10 codes (NBER, 2024).

²<https://icd.who.int/browse10/2019/en>

For each of the above six conditions, patients were split into ‘positive’ vs ‘negative’ cases, based on whether they were given the respective diagnosis at ICU discharge or not, as obtained from their ICD-10 codes. [Table 11](#) shows the resulting distribution of these positive and negative labels across each of the test modalities considered in this study. Note that the total number of labels under the LVD column equals the total number of patients making up the study cohort; this is because all LVD observations are compiled as a single LVD set per patient. Conversely, the total number of labels under X-rays, ECGs and Text exceeds the cohort number, since multiple observations were possible per patient (e.g. each subject could have had multiple X-rays taken during their ICU admission, each counted individually). The Fusion column denotes cases where all 4 types of observations were present for the subjects in question.

The primary evaluation metric was the area under the receiver operating characteristic curve (AUC), which measures the ability of the model to discriminate between classes across all possible thresholds. For threshold-dependent assessments, a threshold of 0.5 was used, reflecting the standard practice when no prior bias toward either class exists. We report AUC as the main metric as it is threshold independent and highlight False Negative Rates (FNR) because missed conditions can cause serious harm; equally, we recognise that False Positive Rates (FPR) are also clinically important in the ICU, where false alarms may lead to unnecessary and sometimes invasive interventions and where alarm fatigue is a serious concern ([Lewandowska et al., 2020](#)). Therefore, thresholds, error rates and the choice of aggregator should be defined based on the condition and context, to balance the risks of both FNR and FPR, with guidance from clinical experts. To improve model robustness, models were trained on balanced datasets using random over-sampling of the training data, after initial experiments showed improved results on validation data. The reported results across the paper were always reported on the original distribution to mimic real world settings. Five fold cross-validation was used for the experiments, with all splits performed at the patient level to prevent data leakage between folds; we ensured that the test data were never seen during any training step, especially when different samples were used for pre-training and fine-tuning. In each fold,

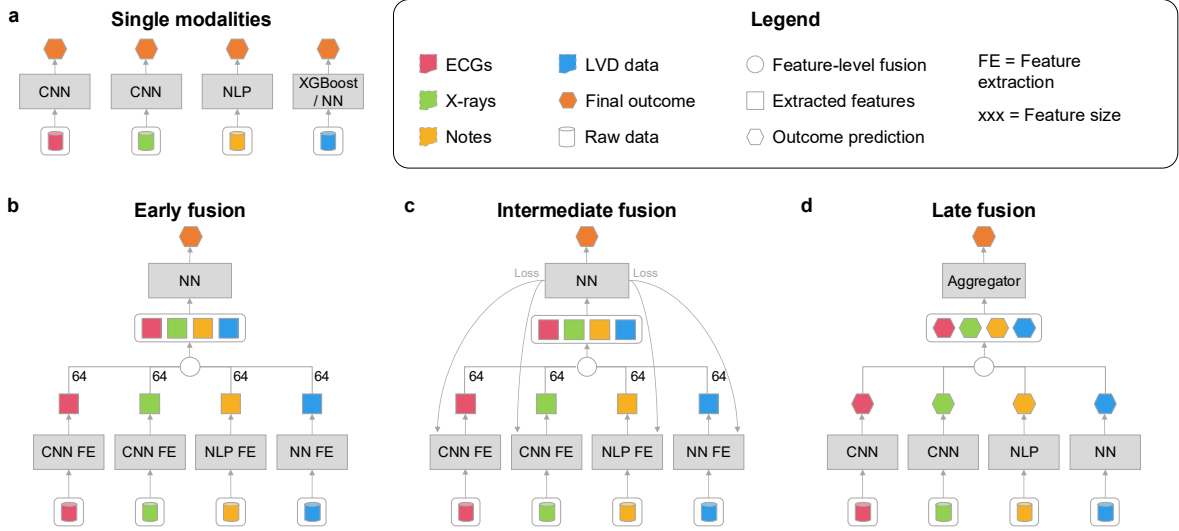


Figure 16: **Fusion model architectures.** Overview of the four fusion model architectures developed and compared in this study. (a) Single-modality models (no fusion); (b) Early fusion, where features from each pre-trained modality are concatenated and jointly processed; (c) Intermediate fusion, with end-to-end training and loss backpropagation across all modalities; (d) Late fusion, where predictions from individual modalities are combined using different aggregation strategies (*average*, *majority voting* and *maximum probability*).

80% of patients were assigned to the training partition and the remaining 20% to the test partition. For training, the training set was further split into training and validation data using an 80/20 split, where the validation set was used for monitoring convergence and learning rate scheduling via `ReduceLROnPlateau`. This resulted in an effective split of approximately 64%/16%/20% for training, validation and testing, respectively. The same patient-level splits were used across all modalities and all experiments to ensure comparable evaluation.

5.2.6 Single-modality models

To establish a benchmark and better understand the contribution of individual modalities, single-modality models were trained first. The following models were developed for each of the four main data types available in MIMIC-IV: low-frequency tabular data (LVD, such as demographics and clinical measurements), high-frequency time-series data or waveforms (ECGs), medical imaging (chest X-rays) and clinical text notes.

Image models. For the image classification of the chest X-rays, a ResNet50 model (He

et al., 2016) was used, following the approach described in Duvieusart et al. (2022). The final classification layer was removed to allow integration into early and intermediate multimodal fusion strategies. Given evidence that convolutional neural networks (such as ResNet50) are less sensitive to weight initialisation in medical imaging tasks than transformer-based models (Ma et al., 2022), pretrained ImageNet weights (Deng et al., 2009) were used, obtained via the `torchvision` package.

Time-series models. High-frequency time-series data or waveforms, in particular ECG recordings. As all ECGs were consistent in sampling frequency and duration, they were processed using spectrograms as proposed in Krones et al. (2023). The recordings were transformed into spectrograms using the `librosa` package and subsequently modelled using the same ResNet50 architecture as applied to image data. This approach was selected as it provides a robust and generalisable alternative to more specialised analyses, achieving up to 95% of the performance of highly tailored methods (Krones et al., 2023).

Text models. Two models were evaluated for clinical text classification. First, ClinicalBERT (Wang et al., 2023), a BERT-based model pretrained on clinical notes, was fine-tuned using the original tokenizer and model weights from HuggingFace³, as described in Sheetrit et al. (2023). Second, OpenAI’s `gpt-4.1-mini-2025-04-14` model (OpenAI, 2024) was used in a zero-shot fashion.

Tabular models. Two modelling approaches were compared for the tabular data containing lab tests, vital signs and demographics: XGBoost and a feed-forward neural network similar to that in Duvieusart et al. (2022). The neural network consisted of three fully connected layers with batch normalisation between layers. The same neural network structure was used for all fusion experiments to ensure better comparability between the fusion techniques. As late fusion can also accommodate XGBoost models and due to the faster training times, we used XGBoost in several ablation studies; these instances are explicitly mentioned.

³<https://huggingface.co/medicalai/ClinicalBERT>

5.2.7 Data fusion approaches

Figure 16 illustrates the three main data fusion strategies applied in this study: early, intermediate and late fusion (implemented as per Krones et al. (2025)). The schematic diagrams in Figure 16 show how information from different modalities is integrated at each stage. In all cases, modality-specific feature extractors (a ResNet50 CNN for X-rays and ECG spectrograms, ClinicalBERT for text, and a feed-forward neural network or XGBoost for tabular data) first process the raw inputs independently. *Early fusion* concatenates the extracted features from each modality and processes them jointly through a shared classifier, without further adaptation of the feature extraction layers. Specifically, each modality-specific model outputs a 64-dimensional feature vector; these four vectors are concatenated into a single 256-dimensional vector, which is then passed through a multi-layer classifier consisting of fully connected layers with batch normalisation (Figure 16b). *Intermediate fusion* differs by allowing error backpropagation through the entire network, which enables optimisation of feature extraction but limits the models that can be used, as the error must be able to be backpropagated. As in early fusion, the modality features are concatenated and passed through the same shared classifier, but here the gradients flow back through all modality-specific feature extractors during training, jointly optimising both the feature representations and the classifier (Figure 16c). This requires all component models to be differentiable, which excludes XGBoost from this setting. *Late fusion* aggregates independent modality-specific predictions using strategies such as average, majority voting, or max, without any additional learned parameters (Figure 16d). Each single-modality model is trained independently and produces a probabilistic output; these outputs are then combined at the decision level. For average aggregation, the output probabilities of the individual classifiers are averaged. In majority voting, the final prediction is determined by the most frequently predicted class after thresholding. The max strategy selects the class with the highest individual prediction probability. Thresholding was applied *after* aggregation for average and maximum. For majority voting, thresholding occurs *before* aggregation to obtain binary votes.

In this work, for early and late fusion, each modality-specific model was trained separately. In the early and intermediate fusion settings, the final classification layer of each single-modality model was removed. For early fusion, the newly added classifier was trained for 30 epochs using fixed features. In contrast, for intermediate fusion, the entire network was jointly optimised for 30 epochs. 30 epochs were chosen based on preliminary experiments that indicated stable convergence of training and validation losses.

In both early and intermediate fusion, the features were combined via a fully connected layer that connected each output node from each modality-specific model to all input nodes of the final classifier. This corresponds to a concatenation of all extracted features into a single vector. To prevent any single modality from having a disproportionate influence, the final layers of all individual models were configured to contain 64 nodes. To ensure that the choice of architecture for early fusion was not suboptimal, we conducted a comprehensive sensitivity analysis. Specifically, we varied both the number of nodes in the final layers of the individual models (between 16 and 128) and the number of layers within the final layers, as all models contained a series of fully connected layers as their final layers.

Model comparison. The performance of the best-performing fusion approach (late fusion, as shown in [Figure 16](#)) was compared with three recent multimodal approaches for healthcare applications: *HAIM* ([Soenksen et al., 2022](#)), *MedFuse* ([Hayat et al., 2022](#)) and *AutoPrognosis-M* ([Imrie et al., 2025](#)). The approaches were selected based on their applicability to the MIMIC-IV dataset, public availability of code, recency (published within the past four years) and relevance to multimodal modelling. *HAIM* ([Soenksen et al., 2022](#)) first creates individual patient files, then generates standardised embeddings from multimodal data and finally applies simple classifiers such as XGBoost. *MedFuse* ([Hayat et al., 2022](#)) employs an LSTM-based fusion module that sequentially integrates clinical time-series and imaging data while handling missing modalities. *AutoPrognosis-M* ([Imrie et al., 2025](#)) automates the construction of multimodal ensembles using AutoML to combine tabular and imaging data via early, joint and late fusion strategies.

Official implementations as published on GitHub were used to ensure reproducibility. The same pre-processing steps as in the original publications were applied (e.g. image resizing). For each baseline, the hyperparameter configuration published in the original repository was retained, including model-specific learning rates, optimisers, schedulers, batch sizes, and any built-in early-stopping or best-checkpoint selection mechanisms. The only common change applied across baselines and our own models was a maximum training budget of 30 epochs, chosen based on convergence checks of our own fusion model. The exact configurations are visible in our public forks linked below. For models not designed to handle all modalities, only the corresponding modality data were used. Where multiple fusion approaches were proposed, the best performing configuration was selected. All fusion experiments, including training and testing under cross-validation, were performed on the same cohort of 5,280 patients that had information available across all modalities (Table 11), so that the comparison between fusion strategies is not confounded by modality-level imputation.

The AutoPrognosis-M model (Imrie et al., 2025) was the only repository not built on the MIMIC-IV dataset; we therefore used the same tabular features and pre-processing steps as in this work. The other models were adjusted to run with Version 3.0 of the MIMIC-IV dataset to ensure comparability. We published the adjustments as forks of the original repositories so readers can track the changes easily. For AutoPrognosis-M (Imrie et al., 2025) this includes the adjustments to run with MIMIC-IV ( GitHub⁴), while we adjusted the MedFuse model (Hayat et al., 2022) to incorporate data from MIMIC-IV Version 3.0 ( GitHub⁵). For HAIM (Soenksen et al., 2022), besides the updates to Version 3.0, several further adjustments were necessary before the code could run successfully ( GitHub⁶).

⁴<https://github.com/felixkrones/AutoPrognosis-Multimodal>

⁵<https://github.com/felixkrones/MedFuse>

⁶<https://github.com/felixkrones/HAIM>

5.2.8 Ablation studies

After determining the most effective single-modality and fusion approaches, three experiments were designed to assess the contribution of individual modalities. First, the fusion model was evaluated using every modality combination to assess the contribution of each modality and determine which combinations most strongly influenced model performance across different conditions. Second, false-negative and false-positive predictions of the fusion network were analysed to determine how often one, two, or all single-modality predictions were incorrect. This analysis aimed to identify whether prediction errors primarily stemmed from weak single-modality models, the fusion strategy itself, or from erroneous single-modality predictions. Third, the behaviour of the fusion model was examined in instances where the best-performing single-modality model produced incorrect predictions. This experiment aimed to assess whether the fusion model would replicate these errors or potentially correct them.

Next, to understand which specific features contributed to the predictions of the model across conditions, the influence of individual LVD features on model performance was assessed using XGBoost feature importance scores. These scores reflect the average gain across all decision trees in the loss function of the model when a feature is used for splitting, thus quantifying how much each feature contributes to the predictive power of the model. Moreover, Shapley values were computed using the Python package `shap` (Lundberg and Lee, 2017) to capture the contribution of each feature to individual predictions.

Given the limitations of clinical resources, a further analysis investigated whether model confidence could guide the need for collecting additional modalities. The idea was that if model confidence is informative, it could not only indicate when additional information would be beneficial but also mimic clinical reasoning, where additional tests are requested mainly when there is uncertainty. This approach was inspired by the Oracle metrics framework proposed by Tran et al. (2022) and the performance gain analysis introduced by Imrie et al. (2025). In this work, the absolute distance

between the probabilistic prediction value and the 0.5 decision threshold was used as the confidence measure. Predictions from single-modality models were ordered by confidence and the corresponding increase in accuracy from applying the fusion model was quantified by calculating the F1-score difference between the single modality models and the multimodal versions.

5.2.9 Prediction improvement over time as data become available

The evolution of prediction quality and modality contributions over time was examined. Based on the typical post-admission timelines for each data type, the analyses investigated when reliable predictions become possible and how the timing of measurements affects performance. Given its strong performance in previous experiments, late fusion was selected as the representative strategy for all ablation analyses.

The availability of X-ray, ECG and LVD data over the course of patients' ICU stay was examined. To assess the impact of data availability on prediction performance, models were trained using only LVD features accessible at successive time points after admission, while keeping the diagnosis at discharge labels as target variables. This analysis focused solely on LVD features, as X-rays and ECG recordings are often already available at the point of ICU admission. The same processing steps as described in [Table 29](#) and [Table 30](#) were used.

5.2.10 Ethical approval

Prior to the start of the project, ethical approval (SSH.OII.CIA.22.108) was obtained through the departmental research ethics committee using the CUREC 1A form, following the University of Oxford's Central University Research Ethics Committee (CUREC) guidelines. Although the study relied exclusively on secondary data analysis of the publicly available, de-identified MIMIC-IV database and did not require additional ethical approval, this step reflects a commitment to responsible research conduct. Access to MIMIC-IV was granted after completion of the required training on research with human participants and acceptance of PhysioNet's Data Use Agreement, which mandates

safeguarding the data, refraining from any reidentification attempts and adhering to data security standards.

5.2.11 Implementation details

For the neural network models a Binary Cross Entropy loss was used and the Adam optimiser (Kingma and Ba, 2014). All images were resized to a resolution of 224×224 , using random cropping, rotation and flipping during training. A learning rate of 0.0001 was set and ReduceLROnPlateau was used as the learning rate scheduler with a minimum rate of 0. All models were fine-tuned for 30 epochs. We used over-sampling during training to correct for imbalances. For fine-tuning either a single NVIDIA L40S or NVIDIA A100 80GB PCIe GPU was used. The term significant denotes results with a p-value of less than 0.05. All statistical tests were two-sided. To assess performance differences between models evaluated on identical cross-validation folds, we used paired-samples t-tests on fold-level performance metrics, thereby accounting for the dependence induced by shared data splits. Data partitioning was performed at the patient level, so that no patient appeared in more than one fold and each patient contributed a single prediction per condition; there were therefore no repeated within-patient observations or cross-fold patient dependencies for which a mixed-effects model would be required. As the low-frequency signals were aggregated into static features rather than modelled as sequences, feature-level autocorrelation did not enter the model comparison. Given the number of pairwise comparisons across fusion strategies, conditions, and baseline models, the resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure, which ranks the raw p-values from smallest to largest and rescales each one by the total number of tests divided by its rank, controlling the expected proportion of false positives among the comparisons declared significant at a false discovery rate of 0.05. For example, with $n = 10$ tests, a raw p-value of 0.01 ranked second smallest becomes $0.01 \times 10/2 = 0.05$ after adjustment.

Table 12: **Comparison of proposed fusion pipeline to state-of-the-art models.** State-of-the-art models from the literature were trained and evaluated on the proposed classification task of six cardiovascular conditions, based on the ICD-10 groupings presented in Table 26. Reported is the average test AUC over 5-fold CV and over the different conditions. If multiple fusion strategies were proposed in the literature, the best-performing was used. For a fair comparison, the same dataset of size 5,280 patients was used for all models (Table 11). Table 31 contains the per disease performance.

Paper/Model	Modalities	Original task	Fusion strategy	AUC mean (SD)
Our model (baseline)	LVD, X-rays, ECGs	ICD-10 codes	Late (avg.)	.773 (.030)
HAIM (Soenksen et al., 2022)	LVD, X-rays	CXR labels	Early	.749 (.066)
MedFuse (Hayat et al., 2022)	LVD, X-rays	ICD-9 codes	Intermediate	.772 (.029)
AutoPrognosis-M (Imrie et al., 2025)	LVD, X-rays	Skin lesion	Late/Ensemble	.763 (.041)

5.3 Results

5.3.1 Participant characteristics

Table 11 shows the prevalence of clinical conditions (as per the diagnosis at the time of discharge) among individuals in the cohort. While some conditions, such as (1) Hypertensive diseases, were present in over half the cohort, others, such as (4) Cardiomyopathies diseases, were diagnosed in only a small proportion. Most static data, including patient demographics, are available immediately after admission, as this is when most background information are recorded and basic vital signs measured. Over time, other data sources, including treatment information, low-frequency measurements, X-rays and ECG signals become available. Table 28 shows patient demographics; the cohort comprises primarily older, white patients with a roughly equal sex distribution.

5.3.2 Model performance

Benchmark comparison Table 12 presents a comparison between the performance of our framework and recent state-of-the-art approaches. The average AUC scores achieved by our model are comparable to the performance of other methods, with no statistically significant pairwise differences (all BH-adjusted $p \geq 0.93$), demonstrating the validity of the proposed approach for the later ablation studies by its ability to perform classification on the target conditions with state-of-the-art performance. The

Table 13: **Condition-wise AUC performance (CV)**. Comparison of the different fusion approaches as shown in Figure 16 per condition. Reported is the AUC. 5-fold CV was used and the average (SD) values are reported. For the multimodal experiments (Figure 16), NN were used for the LVD data. Train set describes whether for training the single modality models all available data were used (‘All’) or only the ones available with all modalities (‘Fusion’). For early and intermediate fusion, the final layers were always trained on the fusion data. For late fusion, the aggregation type used during late fusion is given: ‘avg’ means probabilities were averaged, ‘max’ means the maximum probability was used and for ‘maj’ a majority voting was used. In bold the best performing model with notes and without notes included. Comparing the three fusion strategies without notes (rows with Modalities = All, Notes = No, Train = All), late fusion outperformed both early and intermediate fusion at BH-adjusted $p < 0.05$ in five of the six conditions; the late vs intermediate comparison for chronic ischaemic heart disease was borderline ($p = 0.056$). Early and intermediate fusion did not differ significantly in any condition ($p \geq 0.077$).

Fusion	Modalities	Notes in-cluded	Train set	(1) Hy-perten-sive	(2) Is-chaemic	(3) Chro. is-chaemic	(4) Car-diomyop.	(5) Dys-rhythmias	(6) Heart failure
None	X-ray (CNN)	N/A	All	.741 (.007)	.691 (.019)	.794 (.006)	.721 (.027)	.792 (.012)	.815 (.012)
None	X-ray (CNN)	N/A	Fusion	.675 (.029)	.658 (.035)	.740 (.020)	.637 (.045)	.717 (.013)	.752 (.018)
None	ECG (CNN)	N/A	All	.709 (.010)	.703 (.010)	.751 (.008)	.761 (.011)	.817 (.011)	.808 (.006)
None	ECG (CNN)	N/A	Fusion	.657 (.021)	.561 (.032)	.669 (.014)	.595 (.029)	.726 (.013)	.676 (.022)
None	LVD (XGB)	N/A	All	.798 (.003)	.798 (.007)	.786 (.005)	.738 (.025)	.827 (.005)	.841 (.007)
None	LVD (XGB)	N/A	Fusion	.785 (.014)	.786 (.028)	.784 (.013)	.676 (.039)	.809 (.013)	.810 (.012)
None	LVD (NN)	N/A	All	.791 (.006)	.790 (.008)	.778 (.007)	.727 (.016)	.800 (.004)	.845 (.009)
None	LVD (NN)	N/A	Fusion	.753 (.011)	.746 (.012)	.759 (.012)	.688 (.022)	.764 (.021)	.807 (.020)
None	Notes	Yes (Bert)	Fusion	.860 (.007)	.859 (.013)	.902 (.008)	.783 (.074)	.830 (.017)	.879 (.007)
None	Notes	Yes (gpt)	N/A	.873 (.006)	.795 (.013)	.740 (.014)	.828 (.032)	.810 (.007)	.818 (.006)
Early	All	No	All	.740 (.021)	.728 (.017)	.791 (.025)	.742 (.041)	.808 (.015)	.782 (.036)
Early	All	Yes (Bert)	All notes (fusion)	.826 (.021)	.804 (.028)	.889 (.015)	.770 (.078)	.851 (.020)	.857 (.035)
Interm.	All	No	All	.741 (.023)	.714 (.016)	.775 (.039)	.719 (.073)	.819 (.019)	.813 (.013)
Interm.	All	Yes (Bert)	All notes (fusion)	.827 (.010)	.822 (.022)	.897 (.008)	.789 (.060)	.856 (.017)	.865 (.041)
Late	All (max)	No	All	.775 (.005)	.786 (.023)	.814 (.020)	.795 (.037)	.837 (.013)	.846 (.009)
Late	All (maj)	No	All	.708 (.006)	.682 (.024)	.725 (.012)	.668 (.026)	.744 (.020)	.763 (.011)
Late	All (avg)	No	All	.799 (.009)	.799 (.020)	.824 (.016)	.809 (.032)	.845 (.013)	.860 (.007)
Late	All (avg)	No	Fusion	.757 (.012)	.767 (.016)	.784 (.012)	.732 (.033)	.792 (.009)	.828 (.007)
Late	All (avg)	Yes (Bert)	All notes (fusion)	.863 (.008)	.862 (.015)	.911 (.016)	.846 (.044)	.878 (.012)	.898 (.008)
Late	All (avg)	Yes (gpt)	All	.906 (.009)	.855 (.011)	.871 (.012)	.889 (.025)	.896 (.009)	.916 (.010)
Late	All (avg)	Yes (Bert)	Fusion	.844 (.015)	.843 (.017)	.896 (.011)	.803 (.041)	.854 (.010)	.887 (.012)

table is intended as a sanity check rather than a competitive benchmark; we do not claim superiority over these baselines, and the focus of this chapter remains the systematic comparison of fusion strategies under controlled conditions on the same cohort.

Table 14: **Sensitivity analysis on early fusion architecture.** Condition-wise AUC performance for different variations of late and early fusion, all without clinical notes (see Figure 16). Reported is the macro AUC per condition. 5-fold CV was used and the average (SD) values are reported. Average aggregation was used for late fusion.

Fusion	Train set	Variation	(1) Hyper-tensive	(2) Ischaemic	(3) Chronic ischaemic	(4) Cardiomyop.	(5) Dysrhythmias	(6) Heart failure
Early	All	With single modality models trained on all available data	.740 (.021)	.728 (.017)	.791 (.025)	.742 (.041)	.808 (.015)	.782 (.036)
Early	Fusion	With single modality models trained on available fusion data	.689 (.022)	.657 (.023)	.731 (.033)	.636 (.052)	.722 (.031)	.786 (.014)
Early	All	Reduced final feature layer from 64 to 16 output neurons	.668 (.072)	.693 (.038)	.746 (.042)	.693 (.075)	.758 (.012)	.732 (.040)
Early	All	Increased final feature layer from 64 to 128 output neurons	.712 (.049)	.722 (.026)	.789 (.016)	.719 (.034)	.820 (.021)	.793 (.042)
Early	All	Only one final linear layer instead multilayer network	.746 (.016)	.760 (.011)	.782 (.023)	.776 (.045)	.798 (.0140)	.822 (.017)
Early	Fusion	Only one final linear layer instead multilayer network	.693 (.011)	.728 (.016)	.733 (.017)	.668 (.029)	.722 (.016)	.767 (.020)
Late	All	With single modality models trained on all available data	.799 (.009)	.799 (.020)	.824 (.016)	.809 (.032)	.845 (.013)	.860 (.007)
Late	Fusion	With single modality models trained on available fusion data	.757 (.012)	.767 (.016)	.784 (.012)	.732 (.033)	.792 (.009)	.828 (.007)

Model performance across medical conditions Table 13 shows the model performance for each medical condition. There are substantial differences in both the overall predictability of conditions and the relative performance of different models. Examining single-modality performance, the ability of each modality to predict a condition varies depending on the condition type.

Single modality model performance As shown in Table 13, clinical notes were in all cases the best predictor. However, the results need to be interpreted with care, as the exact timing of the notes is unclear and target leakage cannot be completely ruled out. We therefore treat note-augmented rows as a separate supplementary analysis: the comparison of fusion strategies, the comparison with state-of-the-art baselines (Table 12), and the sensitivity analysis (Table 14) are all based on the no-notes rows of Table 13, so that none of our claims about fusion-strategy robustness depends on the note modality. As shown in Table 13, ECGs were especially helpful in predicting (4) Cardiomyopathies diseases, (5) Dysrhythmias diseases and (6) Heart failure. For the other conditions either the X-rays or LVD data were most helpful. While for (3) Chronic ischaemic heart disease, (4) Cardiomyopathies, (5) Dysrhythmias and (6) Heart failure X-rays and the

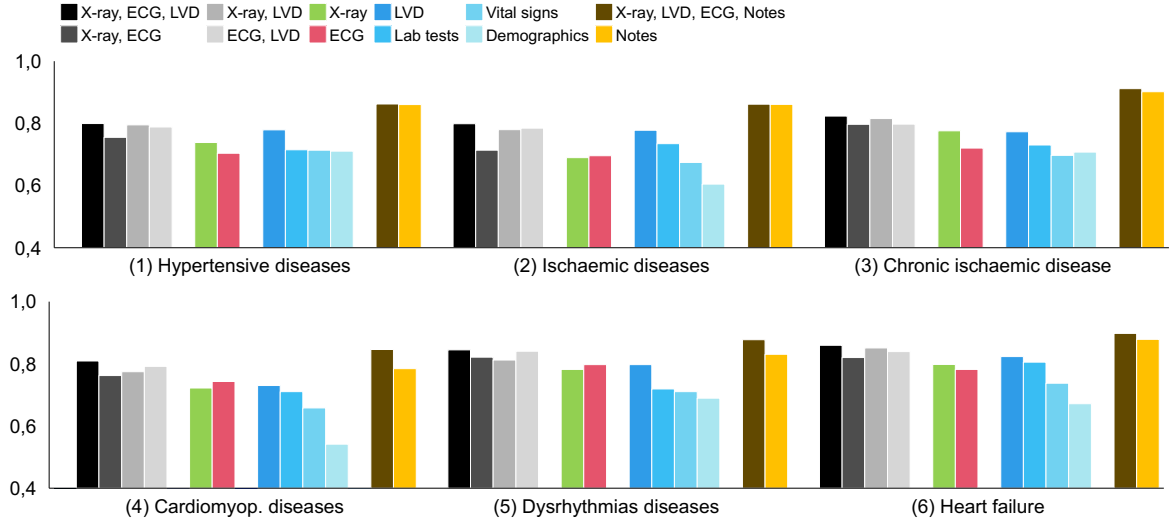


Figure 17: **Effect of modality combinations.** The figure shows the AUC for different combinations of modalities using the late fusion approach, evaluated on the fusion test split. Each plot relates to one of the six investigated conditions (see Table 11). In blue are the different tabular LVD features. The figure shows that only the combination of the different tabular LVD features results in the often strong LVD performance.

LVD data had similar performances, (1) Hypertensive diseases and (2) Ischaemic heart diseases could mainly be predicted by the LVD data.

Multi modality performance As shown in Table 13, the best performing model was usually the late fusion one (as also confirmed by the early-fusion sensitivity analysis in Table 14). For late fusion, the three aggregation operators differ in their performance. The *average* of per-modality probabilities ($p_{\text{avg}} = \frac{1}{M} \sum_{m=1}^M p_m$) acts as a soft consensus: a single extreme p_m shifts the fused score by at most $1/M$, so agreement across modalities drives discrimination. The *maximum* ($p_{\text{max}} = \max_m p_m$) behaves like a permissive “OR”: one confident modality can yield a high fused score, which tends to improve sensitivity (true positive rate) but often increases false positive rate (FPR). Majority voting requires first binarising each modality and then taking the mode; this ignores the magnitude of probabilities, discards information about how confident each modality is and treats all modalities equally regardless of reliability, which together reduce its ability to distinguish between positive and negative cases.

Table 15: **Confusion matrix of the best-performing model.** Counts for the late fusion model (average aggregation, decision threshold 0.5), aggregated across the six cardiovascular conditions (Table 11) and the five cross-validation test folds, so that each patient contributes one prediction per condition ($n = 31,548$). The per-condition matrices are shown in Table 32.

		Predicted	
		Negative	Positive
Actual	Negative	18,315	3,669
	Positive	2,911	6,653

5.3.3 Improvement characteristics

Multimodal effects Figure 17 shows that combining two modalities consistently outperforms the corresponding single-modality models for the six conditions of the circulatory system. However, the magnitude of this effect depends on how similar the individual performances of the two modalities are. If their performances are close, combining them leads to a substantial improvement, as seen for LVD and ECG data in (4) and (5). In contrast, if there is a large performance gap, adding the weaker modality has little effect on the stronger one, as in (1) and (2), where adding ECGs to LVD data yields only minor gains. There is a strong negative correlation ($r = -0.869$, $p < 0.001$) between the difference of the combined predictors and the resulting improvement over the stronger predictor. This means that, on average, smaller differences are associated with larger improvements. The correlation was calculated across all diseases and all possible predictor combinations, including those involving two or more predictors. The benefit of including a third modality depends on whether it compensates for a previously missing strong modality. For instance, when two strong modalities – such as X-rays and LVD data in case (1) on Hypertensive diseases – have already been combined, adding another modality (i.e., ECGs) provides little additional benefit.

Figure 32 shows cases where the main single-modality predictor would have been incorrect but was corrected by the fusion model. In most cases, the multimodal model fixed the false positives cases, except for (1) Hypertensive diseases, which is also the only condition with substantial more positive cases in the training data. Here, the multimodal

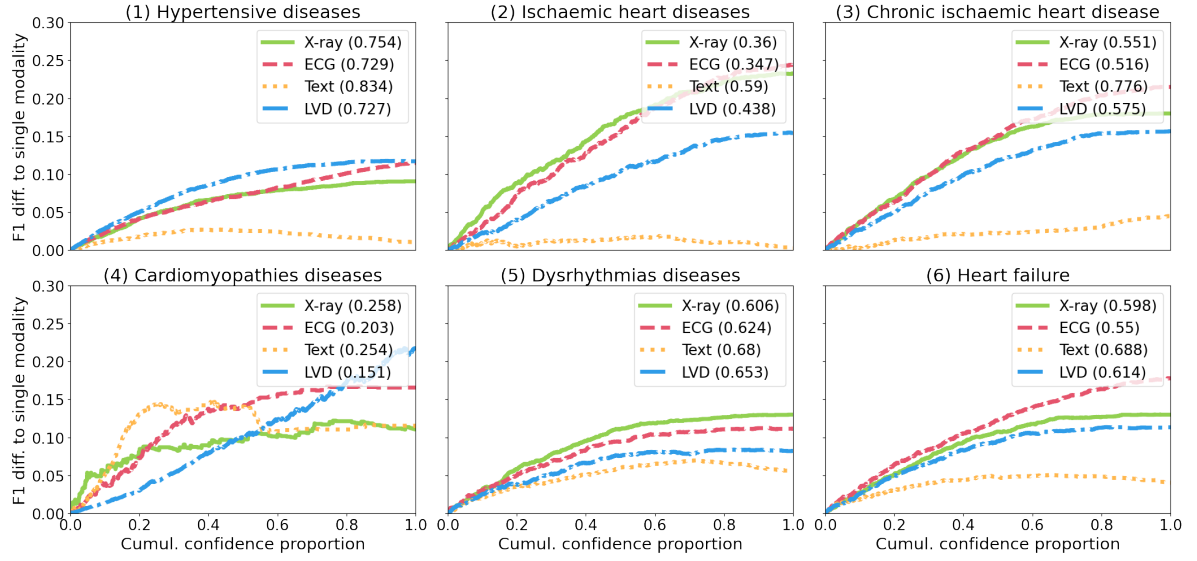


Figure 18: **Selective acquisition of multimodal data.** The plot shows the cumulative F1-score improvement of the multimodal model over the corresponding single-modality models. The x-axis represents the cumulative proportion of patients for which the multimodal was used instead of the single modality model, ranked by the certainty of the single modality predictions, whereas the y-axis represents the difference in F1-score for that scenario, compared to the ‘base’ F1-score of the single modality (which is the value indicated in parentheses in the legend for each modality). For example, at 20%, the 20% most uncertain predictions of, e.g. the X-ray model were replaced with the multimodal predictions. Then the F1-score was calculated over all predictions and compared with the F1-score for just the X-rays. The graph indicates that, as the certainty of single-modality predictions increases (towards the right of the x-axis), the added value of the fusion model diminishes (the graph flattens out). This pattern suggests that the multimodal approach is most advantageous when single-modality predictions are less confident. Each plot relates to one of the six investigated conditions (see Table 11). Figure 34 shows the F1-score improvement per ten percent bin.

approach mainly fixed the false negative cases.

To make the real-world implications of these errors explicit, Table 15 reports the confusion matrix of the best-performing model, late fusion with average aggregation, aggregated across the six conditions and the cross-validation test folds. Of the 9,564 positive cases, the model missed 2,911 (false negatives), while it produced 3,669 false positives among the 21,984 negative cases. The per-condition confusion matrices are given in Table 32. They show that the error profile depends strongly on prevalence, for example the low number of correctly identified cases for the rare condition (4) Cardiomyopathies diseases.

Figure 33 shows two main findings: First, models with a higher FNR usually have lower FPRs (compare numbers above bars in i) and iii)). However, the relationship is

not linear, which means that one can tweak the model into one or the other direction depending on which error leads to the higher costs / is more severe. As long as the class imbalance is not very extreme, as, for example, for (4) Cardiomyopathies diseases (with only 5% positive cases), the class imbalance is not directly correlated to the error rates. For example, (1) Hypertensive diseases and (3) Chronic ischaemic heart diseases show similar error rates, even though (1) has almost two-third positive cases while (3) has over two-third negative cases. Second, [Figure 33](#) shows that in most of the cases where the fusion model was wrong, either all or at least more than half of the single modality models were wrong. However, the figure also shows that in up to 17% of the cases, a single incorrect prediction from one modality caused the fusion model to produce an incorrect result, even though the other two single-modality models would have been correct. As *average* aggregation of positive class probabilities was used with a fixed decision threshold, a strongly confident but incorrect modality can override two correct but less confident predictors, where confidence is measured as distance to the decision threshold.

Data efficiency [Figure 18](#) and [Figure 34](#) examine when adding extra modalities is most beneficial and [Figure 35](#) shows a deep-dive analysis into the probability distributions of the different models. As the figures show, the impact on model performance of adding additional modalities decreases as the algorithmic certainty of single-modality models increases. In clinical practice, this could suggest first using the most suitable single modality for a given condition and considering further tests primarily in cases where the model indicates high uncertainty. However, it is important to distinguish algorithmic uncertainty from clinical uncertainty, as low algorithmic confidence does not necessarily imply a clinically ambiguous case. Exploring this more formally, for example by investigating different measures of model uncertainty and implementing system-wide human-in-the-loop approaches where uncertain predictions are referred to a clinician at a defined cost in the decision process, represents a promising direction for future work. This highlights the importance of communicating not only absolute predictions but also model certainty ([Tran et al., 2022](#)). Providing such information when deciding whether

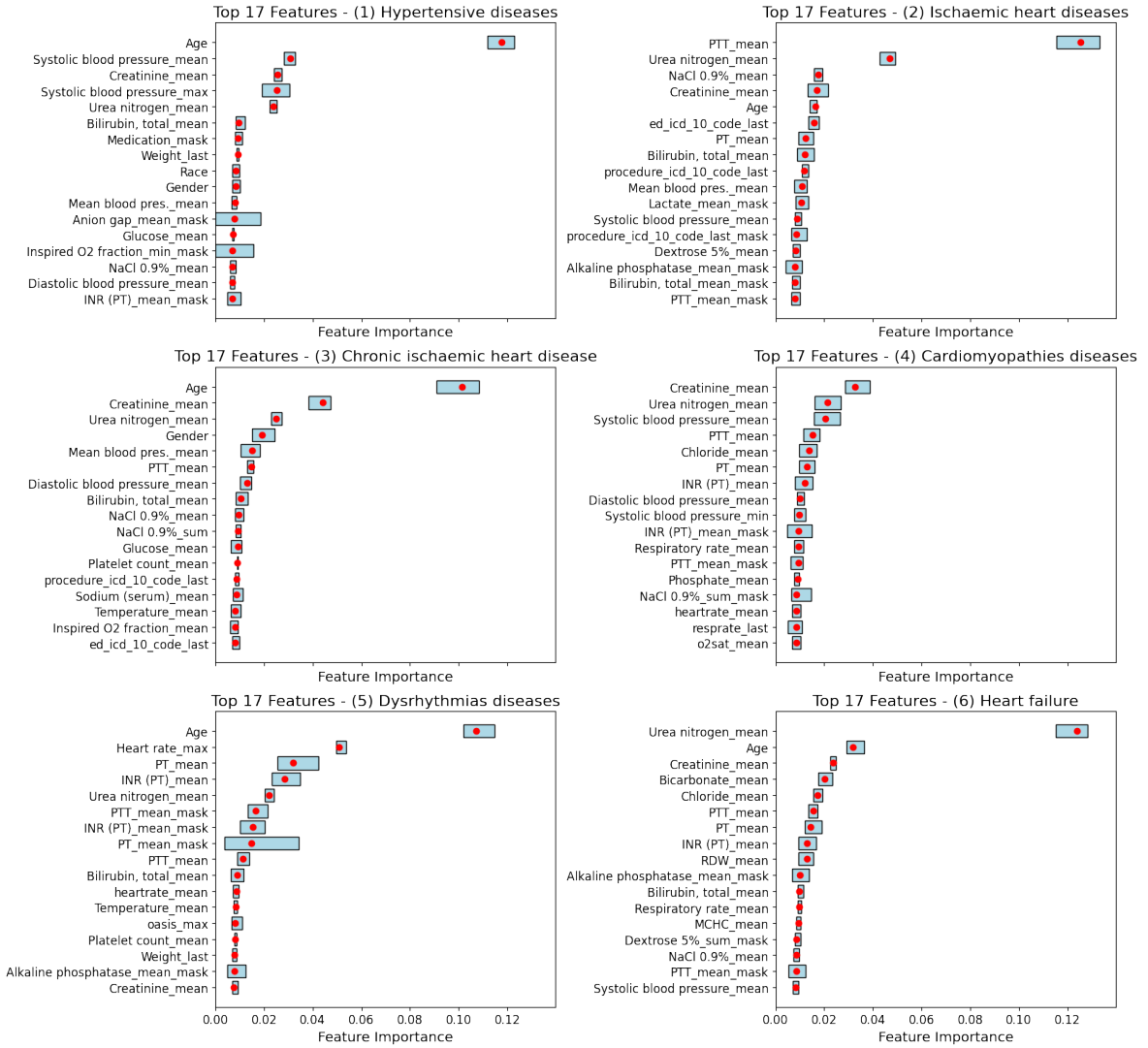


Figure 19: **Feature importance of LVD data.** Top features of the LVD XGB models for the various conditions, shown as their contribution to classification performance. Plotted is the range from minimum to maximum importance across folds, ordered by the mean importance (red dots). Each plot relates to one of the six investigated conditions (see Table 11). These features were found to be informative for prediction, but are not necessarily causal factors.

to collect additional data can (a) reduce the risk of incorrect predictions and (b) lead to more cost-effective workflows.

Feature effects Figure 19 shows that for most diseases, a patient's age is among the top five most important features for classification, with higher age associated with a higher disease probability (Figure 20), except for cardiomyopathies, where no demographic

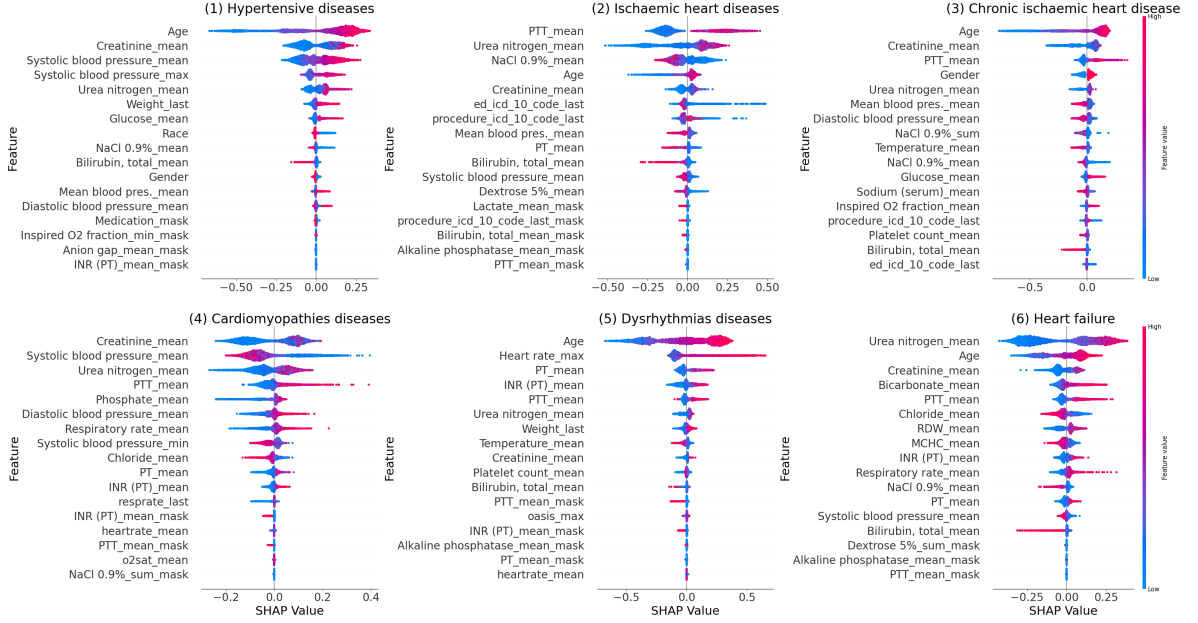


Figure 20: **Shapley values for the top most important LVD features.** Plotted is the distribution of Shapley values from the XGBoost model across all samples, where each point represents one patient, the position on the x -axis indicates the impact on the model output and the colour encodes the feature value, with higher values shown in red and lower values in blue.

information appears in the top ten. Blood pressure, PTT⁷ and urea nitrogen⁸ are among the most informative features for most diseases, where “informative” refers to variables that contribute to the classification decisions of the model rather than to direct causal factors. As shown in Figure 17, the predictive strength of the LVD data arises from the combination of different feature types, as the individual types perform significantly worse.

As mentioned above, we only included features that were available for at least a third of the patients. However, based on clinical guidance, we also examined where Troponin and Brain Natriuretic Peptide (BNP) would rank (Lyngbakken et al., 2019). Using a subgroup of 822 patients with both features available (80% used for training), we re-trained the XGBoost model with these features included. As expected, Troponin ranked among the top ten most important features for (1) Hypertensive diseases, (2) Ischaemic heart diseases, (3) Chronic ischaemic heart diseases and (6) Heart failure, while

⁷PTT (Prothrombin time) is a blood test that measures the time it takes for blood to clot, helping assess coagulation disorders and monitor heparin therapy.

⁸Urea nitrogen is a waste product formed from protein metabolism, measured in blood to evaluate kidney function and hydration status.

being the top predictor for (2) Ischaemic heart diseases and (3) Chronic ischaemic heart diseases. BNP ranked among the top ten most important features for (1) Hypertensive diseases, (3) Chronic ischaemic heart diseases, (5) Dysrhythmias diseases and (6) Heart failure, while being the top predictor for (6) Heart failure.

5.3.4 Prediction improvement over time as data become available

Figure 21 shows that model performance for LVD data improves, particularly during the first hours of the ICU stay, as most data become available in this period. Figure 22 indicates that this improvement is strongly related to the timing of lab tests, especially for conditions that heavily rely on such tests (e.g. (4) Cardiomyopathies). Although most lab tests are conducted within the first few hours of an ICU stay or even earlier, the figure also reveals that some of the most predictive features are recorded only for a subset of patients during this time. While performing these tests earlier could improve predictive performance even prior to ICU admission, test ordering is typically guided by clinical suspicion and requires balancing potential diagnostic benefit against costs and risks. In this context, the framework could be extended to simulate hypothetical positive versus negative findings, allowing clinicians to assess whether the expected reduction in uncertainty justifies the test. This ‘value of information’ approach would help identify when a specific test offers enough differentiating power to significantly alter the prognosis or narrow the differential diagnosis, thereby guiding more targeted test ordering.

Many information are already available before patients are admitted to the ICU. In particular, X-ray and ECG acquisition is often decoupled from the inpatient stay and must be retrospectively mapped to the corresponding hospital stay. Only vital signs and lab test data show a broader distribution in terms of when they are first collected.

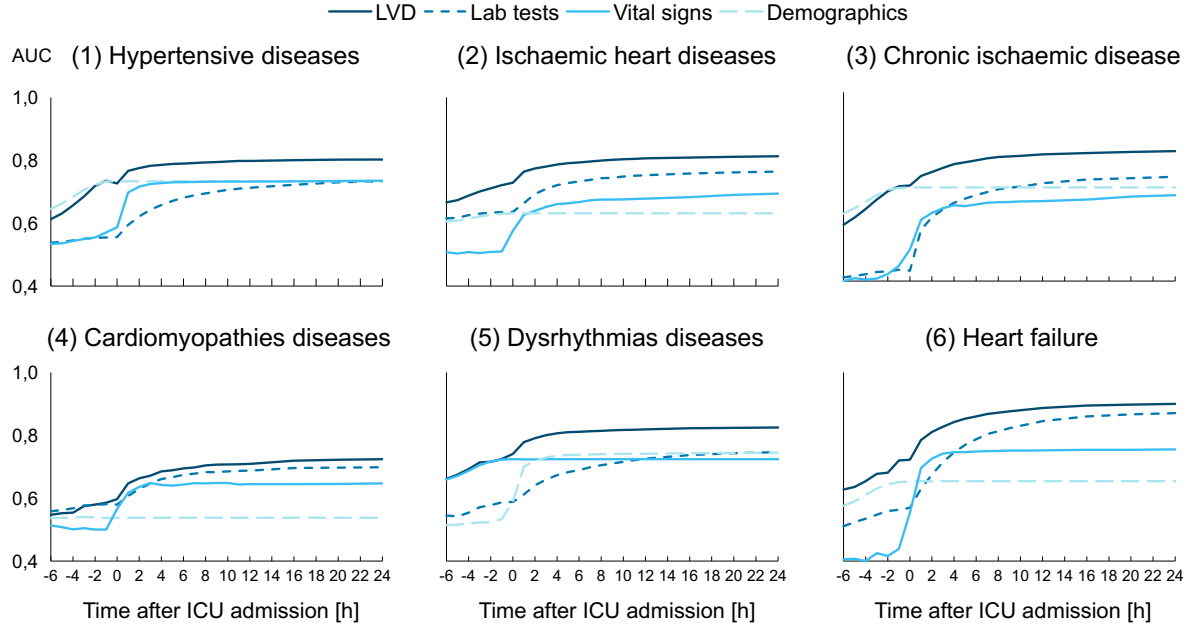


Figure 21: **Improving model performance with increasing data availability.** Performance of the XGBoost model as a function of feature availability over time after ICU admission. Each plot corresponds to one of the six investigated conditions (see Table 11).

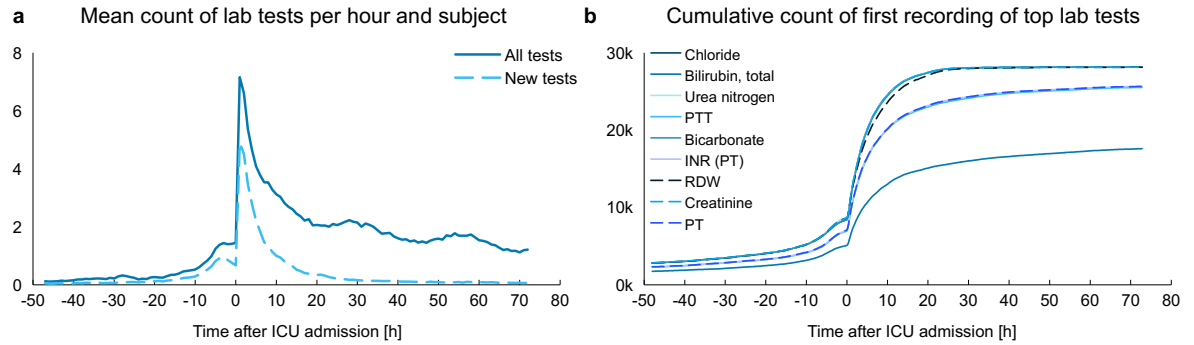


Figure 22: **Availability of lab tests.** *Left:* Average number of lab tests over time: overall (solid blue) compared with counts including only the first occurrence of an event (light blue, dotted). *Right:* Cumulative count of the top lab tests (those ranking among the top 10 features across diseases, see Figure 19) at the time they were first conducted per patient. 55 subjects did not have any lab events and are excluded here.

5.4 Discussion

5.4.1 Main findings

In this work we explored the conditions under which multimodal approaches are most helpful in clinical settings, by comparing and investigating different fusion approaches for integrating multimodal data in the prediction of cardiovascular conditions in ICUs.

By comparing early, intermediate and late fusion strategies, we showed that late fusion offers the most reliable performance under clinical constraints. The models combined imaging, tabular and time-series data, allowing for a comprehensive assessment of how fusion strategies affect predictive accuracy and robustness in ICU settings.

Our primary aim in this study was to evaluate and compare multimodal fusion techniques rather than to advance clinical knowledge about individual biomarkers or to provide concrete test recommendations to clinicians. While we designed the experiments to be as clinically realistic as possible, the translation of these findings into actionable diagnostic guidance, including the evaluation of clinically sparse but highly relevant tests such as for Troponin or BNP, remains an important direction for future work.

Fusion comparison A central finding is the robust effectiveness of late fusion, across the architectures we explored. Across all 18 fusion-strategy comparisons in [Table 13](#), late fusion outperformed both early and intermediate fusion at $p < 0.05$ in five of the six conditions, while early and intermediate fusion did not differ significantly in any condition. While early and intermediate fusion strategies theoretically enable richer feature representations by modelling cross-modal interactions ([Guo et al., 2019](#); [Holste et al., 2021](#); [Soenksen et al., 2022](#)), they often require larger datasets and more complex model architectures ([Huang et al., 2020b](#)), which are frequently impractical in ICU applications. Late fusion, by contrast, achieved high predictive performance with lower data requirements and simpler model design, making it more suitable for clinical implementation. As shown in our sensitivity analysis ([Table 14](#)), early fusion performance varies with architectural choices, and we cannot rule out that a more extensive architecture search could narrow the gap. Our claim is therefore that late fusion is the most robust and best-performing strategy under the architectures and data constraints considered, rather than universally superior. These results support earlier findings by [Huang et al. \(2020b\)](#) ([Table 2](#)), who also identified late fusion as the most practical strategy and thus contribute to the understanding of optimal fusion strategies in critical care, directly addressing research question **RQ1**: *Does any fusion method*

consistently achieve superior performance in multimodal prediction of cardiovascular conditions in the ICU?

Modality effect The results show that the predictive value of each data modality varies by condition and that combining modalities generally improves performance, as shown in nearly all cases in [Figure 17](#). This supports the application of multimodal learning to clinical prediction tasks, particularly when the modalities provide complementary information. However, the findings also reveal two important caveats. First, when one modality already captures most of the relevant signal, adding further modalities yields only minor gains, which may not justify the additional complexity or cost. Second, in rare cases where an added modality primarily introduces noise, performance may decline. This indicates that additional modalities can impair performance when they fail to contribute meaningful signal. Misclassification analysis further showed that fusion models rarely corrected errors when all single-modality models failed ([Figure 33](#)). The largest gains occurred when combining two modalities of similar predictive strength ([Figure 17](#)). The findings directly address research question **RQ2**: *What is the contribution of each individual data modality to predictive performance for cardiovascular conditions in the ICU?* From a clinical perspective, the findings suggests that combining data sources is generally beneficial and can improve decision support accuracy. At the same time, the findings emphasize the need for careful selection of modalities to avoid unnecessary overhead and the risk of reduced performance in specific cases.

The study also highlighted the relevance of evaluating model confidence ([Tran et al., 2022](#)), particularly when determining whether further data collection is warranted. Providing such information when deciding whether to collect additional data can (a) reduce the risk of incorrect predictions and (b) lead to more cost-effective workflows. Efficiency curve analysis demonstrated that the benefits of multimodal fusion were most evident in cases with low prediction certainty ([Figure 18](#)), consistent with earlier studies on uncertainty-based decision-making ([Imrie et al., 2025](#)). [Figure 18](#) can therefore be read as a cost-versus-benefit curve for a selective acquisition policy: the x-axis is the

proportion of patients for whom the multimodal model is used, ranked from most to least uncertain by their single-modality probability, and the y-axis is the resulting cumulative F1 gain. This suggests a possible framework to help steer the decision if acquiring more data (i.e., ordering more tests) is necessary, in which additional tests are conducted only when the uncertainty of the model exceeds a predefined threshold. For that, the model could be extended to simulate positive and negative inputs to evaluate the differentiability potential of the extra test with respect to the resulting predictions. Such an approach may reduce unnecessary investigations, lower costs and lessen patient burden, while preserving diagnostic accuracy. Moreover, by investigating different measures of model uncertainty and implementing system-wide human-in-the-loop approaches where uncertain predictions are referred to a clinician at a defined cost in the decision process, represents a promising direction for future work. A formal separation of epistemic and aleatoric uncertainty, and a prospective simulation of such a clinician-in-the-loop workflow, would require prospective data or a dedicated uncertainty-modelling study and are left to future work.

Timing of data availability The temporal analysis showed that the timing of data availability plays a central role in determining predictive performance, consistent with previous findings in other clinical contexts (Deasy et al., 2020; Placido et al., 2023; Thiele et al., 2025). X-ray and ECG data are often available before or at the time of ICU admission, while vital signs and laboratory tests are usually collected in a staggered manner during the first hours of the ICU stay (Johnson et al., 2024). Model performance improved significantly within the first 24 hours, particularly when laboratory results became available. This suggests that if laboratory test results were available prior to admission, they could further support early assessment and inform better allocation of ICU resources. These findings directly address **RQ3**: *At what point during a patient’s stay in the ICU has data been sufficiently accumulated to make reliable predictions?* The answer depends on which modalities contribute most to predictive performance. When laboratory values or vital signs are the primary drivers, reliable predictions become possible only after sufficient data are collected during the first 24 hours. In contrast,

when X-rays, ECGs, or demographic features are most informative, accurate predictions could potentially be made earlier. However, as our analysis used the final available imaging and signal data, this potential for early prediction relies on the assumption that pre-admission tests hold comparable diagnostic value to those obtained during the ICU stay. Furthermore, the reliance on accumulated data highlights the value of temporal trends. While our approach implicitly captures some trajectory information by comparing mean and last values, it does not explicitly model rates of change. Given that clinicians often prioritize trends over absolute values, future work could focus on explicit trend features, though this inherently requires a sufficient observation window and recording frequency to derive meaningful gradients. In our cohort, the low-frequency laboratory and vital-sign measures are stored as aggregated values rather than dense time series, so derivative features such as lactate clearance or haemodynamic slopes, and time-aware architectures designed for densely sampled signals, cannot be applied reliably at this resolution.

5.4.2 Strengths of the study in relation with other studies

A key strength of this study is its systematic investigation of multimodal fusion strategies across a broad set of cardiovascular conditions in the ICU, addressing limitations of prior work that typically focused on individual diagnoses or a small set of modalities (Krones et al., 2025). While previous studies have examined fusion strategies in restricted contexts – such as pulmonary embolism detection (Huang et al., 2020c) or dermatological image classification (Imrie et al., 2025) – this work provides a comprehensive comparison of early, intermediate and late fusion across six clinically relevant conditions. By using the MIMIC-IV dataset (Johnson et al., 2024), which integrates imaging, time-series signals, LVD data and clinical text, this study moves beyond earlier approaches that relied on a single dominant modality with limited auxiliary features. The consistent performance of late fusion across tasks offers clarity in light of mixed findings from prior studies, where no fusion strategy showed universal superiority (Guo et al., 2019; Holste et al., 2021; Huang et al., 2020c; Imrie et al., 2025).

One limitation in studies addressing multiple conditions in the ICU so far was how prediction targets were defined. Many relied on radiology reports or clinical notes (Huang et al., 2020c; Soenksen et al., 2022), which are subject to biases introduced during the diagnostic process or by the referring clinician (Duvieusart et al., 2022). In contrast, ICD codes summarise diagnoses across the entire hospital admission and are less susceptible to such biases (Hyland et al., 2023). Although recent work has begun to adopt ICD codes as prediction targets (Lopez Alcaraz and Strodthoff, 2024; Strodthoff et al., 2024), these studies were not fully aligned with the experimental setup proposed here. The limitations were both technical and clinical. Technically, the models were either limited to one fusion approach and did not compare fusion strategies (Lopez Alcaraz and Strodthoff, 2024) or attempted multi-label prediction of more than 1,000 ICD codes (Strodthoff et al., 2024), which can dilute performance, as models tend to disproportionately optimise for the most frequent codes (Wang et al., 2024). Clinically, they used limited observation windows (e.g. only the first 1.5 hours of admission (Lopez Alcaraz and Strodthoff, 2024)) or the models focused on a single modality (Strodthoff et al., 2024). To ensure clinically meaningful groupings and more balanced data distributions, the focus in here was on grouped ICD codes, selecting conditions with similar underlying pathophysiology (see Table 26).

The presented data preparation pipeline offers several advantages over existing approaches used to process the MIMIC-IV dataset. Unlike previous solutions, such as HAIM-MIMIC-MM, MIMIC-IV-Ext-MDS-ED and Symile-MIMIC, this pipeline integrates all major modalities available in MIMIC-IV, including ECG waveforms and chest X-rays and combines them into an event-based format. This comprehensive approach improves temporal alignment across data sources, addressing limitations in earlier efforts, such as the exclusion of ECG data in HAIM-MIMIC-MM (Soenksen et al., 2022) and the restricted imaging data in MIMIC-IV-Ext-MDS-ED (Lopez Alcaraz and Strodthoff, 2024). Another key improvement is the modular and scalable nature of the developed pipeline. While previous approaches relied on rigid inclusion criteria or outdated software requirements (Soenksen et al., 2022), the current framework is designed to

adapt to future versions of MIMIC-IV and to operate independently of local computing environments. This flexibility enhances reproducibility and facilitates the extension of the pipeline to different settings and cohorts. Additionally, the hybrid storage strategy optimises data handling by consolidating low-intensity information while externally referencing high-intensity data. This structure enables efficient loading during model training. In summary, the developed pipeline not only addresses the fragmentation and scalability issues of MIMIC-IV but also establishes a reproducible and comprehensive framework that helps with future multimodal research. This structured approach ensures improved temporal consistency, efficient data management and richer clinical feature representation compared to previous solutions.

5.4.3 Limitations

The models were developed and evaluated solely on the MIMIC-IV database ([Johnson et al., 2024](#)). This may restrict generalisability to other clinical environments with different patient characteristics, data availability and recording practices. Although six cardiovascular conditions were included to broaden the scope, the analysis remains dataset-specific. This reflects an inherent challenge of multimodal research: the limited availability of comprehensive datasets limits the comparisons options across studies and settings. In [Krones et al. \(2025\)](#) we summarised available multimodal datasets.

A further limitation concerns the way predictors were selected. We retained tabular features that were available for at least a third of patients or had been used in prior work, which is a data-driven rule based on recording frequency rather than clinical importance. A purely data-driven selection of this kind can discard markers that are clinically informative but recorded only for a minority of patients, so that the data, rather than the underlying physiology, decide what the model can learn from. The Troponin and BNP analysis illustrates this risk directly: both are clinically established cardiac markers that fell below the availability threshold, yet when we re-introduced them on the subgroup of patients for whom they were measured, they ranked among the most important features and were the single strongest predictor for ischaemic heart

disease and heart failure, respectively. This shows that input from clinicians remains important alongside data-driven pipelines, because expert knowledge can flag variables that an availability-based filter would otherwise remove. It also points to a wider tension: optimising only for what is well recorded can entrench existing measurement practices and miss signals that are rare in the data but central to the condition. Combining data-driven feature selection with clinically guided inclusion of sparse but high-value markers, and treating their low availability as a measurement problem to be addressed rather than a reason for exclusion, is therefore an important direction for future work.

We focused on AUC, complemented by FNR, FPR and F1. Extending this with reliability diagrams, Brier scores and formal Decision Curve Analysis for the late fusion model and its single-modality components, ideally on an external cohort, would be a natural extension of this work before any clinical deployment.

The hospital from which MIMIC data were collected is located in Boston, MA, USA and only data from patients there were collected. Hence, it is limited to the demographic structure of that region. Moreover, because it is a teaching hospital, it can be assumed that the treatment patients receive is rather extensive compared to the world’s average healthcare supply. As a result, several of our quantitative findings, including the 24-hour data-availability window and the relative ranking of modalities, reflect the workflows and ordering practices of this single, well-resourced US academic centre and may differ in resource-limited or non-academic settings. External, multi-site validation is therefore a critical next step before any clinical implementation. In particular, the staggered availability of laboratory tests, vital signs and imaging in resource-limited settings would shift both the optimal time window for prediction and the relative usefulness of individual modalities, an effect that an external-cohort study with different ordering practices would be best placed to quantify. However, as it is a single recording site, the data quality can be assumed to be constant across different demographic groups.

Diagnostic labels were based on ICD codes, which are mainly used for administrative purposes. This may introduce label noise or systematic bias caused by insurance premium

and reimbursement schemes (Dragan et al., 2022). An independent validation of these labels is not possible within MIMIC-IV, since the only available reference, the free-text notes, is itself one of the modelled modalities; label-noise-aware training or probabilistic label treatment is therefore a direction for future work rather than something we can verify on this dataset. Furthermore, patients who died during the hospital stay were excluded. This concerned 4,716 patients (Table 27), a group likely to include some of the most severely ill cases, so the cohort underrepresents the highest-acuity end of the ICU population. While this exclusion could introduce a bias by omitting the most severe cases, it allowed us to ensure consistent data availability across modalities and time points, which was essential for comparing fusion techniques. We did not model mortality as a competing risk or within a joint survival-classification framework because our prediction target is the diagnostic classification of cardiovascular conditions from ICD-10 codes rather than time-to-event or survival; such frameworks would address a different question rather than correct the cohort composition.

A further limitation concerns the comparison with state-of-the-art baselines. We retained the published hyperparameter configurations of HAIM, MedFuse and AutoPrognosis-M rather than performing an independent hyperparameter search per method on our cohort. This choice prioritises reproducibility and stays close to the configurations the original authors reported as best for their settings, but we cannot rule out that a method-specific search on our data would close part of the observed performance gap. The same applies to the choice of baselines: we included the three recent multimodal approaches whose code was publicly available and could be adapted to MIMIC-IV v3.0 within the time constraints of this work, and our published forks make any further re-tuning or extension by other researchers straightforward. A systematic comparison of computational cost, training time and inference latency across fusion strategies was beyond the scope of this study, which focused on predictive performance and the conditions under which fusion is beneficial; we note, however, that late fusion supports parallel execution of per-modality models and allows lightweight, modality-appropriate architectures (for example gradient boosting on tabular data), which can offset its apparent overhead at

inference time. For the averaging aggregator used in this work, the aggregation step itself has negligible inference cost compared with the per-modality forward passes, so the wall-clock latency of late fusion is dominated by the slowest single-modality predictor rather than by the fusion step. For the clinical-text modality, the cleaning pipeline relies on a proprietary API; although we use a pinned model snapshot (`gpt-4.1-mini-2025-04-14`) and release the prompts, exact long-term reproducibility outside the OpenAI API cannot be guaranteed. We also note that, given the way MIMIC-IV stores discharge notes, a temporal-control experiment that predicts the diagnosis from notes written strictly before it was made cannot be implemented cleanly within this dataset.

Finally, only a subset of modalities was included that reflects data sources typically available in real-world settings. Computed tomography, echocardiography and wearable sensors were excluded due to limited data availability. Echocardiography in particular represents an important diagnostic tool in cardiopulmonary care; however, its integration was not feasible in this study. First, echocardiographic video analysis constitutes a separate research domain that requires dedicated pipelines and was therefore beyond the scope of this work. Second, while text-based reports would in principle provide a more direct way of incorporating echocardiography, these notes are not yet publicly released in MIMIC. In practice, extending our late-fusion pipeline to include them would be straightforward. Beyond echocardiography, future extensions should also consider modelling temporal trends over longer time windows and systematically evaluating how varying the inclusion of clinically important tests, such as Troponin or BNP, affects diagnostic differentiation and model uncertainty. A further direction is the systematic evaluation of fusion performance under missing modalities, comparing imputation schemes and quantifying how performance degrades as individual modalities become unavailable across the different conditions; the pipeline already supports masking for this purpose, but a thorough analysis is left to future work.

5.4.4 Study implications

This study showcases the value of multimodal machine learning models in addressing the diagnostic complexity of ICU settings. By systematically evaluating fusion strategies across a broad set of cardiovascular conditions, it demonstrates that combining heterogeneous data sources, such as vital signs, demographics, ECGs, laboratory tests and imaging, can improve predictive performance and support earlier, more reliable clinical decision-making (Cahan et al., 2023; Khader et al., 2023). The comparative analysis of fusion methods revealed that late fusion offers consistent benefits under realistic constraints of data availability, particularly in cases of diagnostic uncertainty. While our analysis focused on ICU data from MIMIC, the framework itself, that is, the systematic comparison of fusion strategies, the modality and timing ablations and the uncertainty-based ranking, is database-agnostic and can be re-applied to the analysis of other multimodal databases, such as the UK Biobank⁹. The specific quantitative findings, however, should not be assumed to transfer without external validation.

Moreover, the findings provide practical guidance for prioritising data collection efforts and integrating machine learning models into clinical workflows (Acosta et al., 2022). For clinicians, the automated integration of multimodal information has the potential to refine risk stratification and optimise intervention timing. For policymakers, the results support initiatives to improve the interoperability of electronic health record systems and to promote infrastructure that enables the effective use of multimodal data. Ultimately, this work advances both methodological understanding and practical readiness for deploying uncertainty-aware, multimodal AI systems in critical care. We stress, however, that the reported gains are improvements in predictive performance and not yet validated diagnostic guidance: the framework is intended to support, not replace, clinician judgement. A path towards clinical use would require external multi-site validation, prospective evaluation in the intended setting, and regulatory assessment. We therefore position this study as a methodological benchmark rather than a deployable end-point, and view emerging clinical foundation models as a complementary direction

⁹<https://www.ukbiobank.ac.uk/>

that a systematic comparison of fusion strategies can inform.

5.4.5 Data and code availability

All data used in this study are derived from the publicly available MIMIC-IV database and its associated modules. Access to MIMIC-IV requires successful completion of the PhysioNet credentialing process, including human subjects research training and acceptance of a data use agreement. More information is available at <https://physionet.org>.

We published the adjustments to all repositories. For AutoPrognosis-M (Imrie et al., 2025) this includes the adjustments to run with MIMIC-IV ( GitHub¹⁰), while we adjusted the MedFuse model (Hayat et al., 2022) to incorporate data from MIMIC-IV Version 3.0 ( GitHub¹¹). For HAIM (Soenksen et al., 2022), besides the updates to Version 3.0, several further adjustments were necessary ( GitHub¹²).

The developed code will be published via PhysioNet, as we have also published previous work (Duvieusart et al., 2023, 2024). In contrast to those projects, which were shared as datasets, this submission will be released as a PhysioNet Software Project, similar to the HAIM repository (Soenksen et al., 2022). Publishing the pipeline as software ensures that the underlying data remain protected under the MIMIC-IV access controls and makes it easier for others to rerun the code if the dataset is updated in the future. The software package will include code for data extraction, preprocessing and integration across MIMIC-IV modalities; scripts to convert the data into an event-based multimodal format; and visualisation tools to support expert exploration. This setup enables credentialed researchers to reproduce the full preprocessing workflow using their own local copy of MIMIC-IV. To support reproducibility, the release specifies the software environment through a pinned requirements file, controls all model and preprocessing settings through a single configuration file, uses a fixed random seed throughout data splitting, resampling and training, and stores the exact patient-level cross-validation fold assignments (a 5-fold scheme with a nested 64/16/20 train, validation and test split) so

¹⁰<https://github.com/felixkrones/AutoPrognosis-Multimodal>

¹¹<https://github.com/felixkrones/MedFuse>

¹²<https://github.com/felixkrones/HAIM>

that all experiments reuse identical splits.

6 Ablation on subgroup differences in multimodal data fusion

Abstract. Machine learning systems are increasingly being developed for high-stakes decision-making across many domains, including healthcare. However, these systems often show subgroup performance differences and group-dependent errors. This risk may be amplified in multimodal fusion, where images are combined with tabular clinical data. Demographic information can be present both explicitly, through tabular inputs, and implicitly, through images, which risks double counting. We hypothesised that this redundancy could induce group-dependent miscalibration. In the first part of this study, we demonstrate that such subgroup differences exist, using cardiovascular condition prediction in ICUs as a case study. We then show that these differences were not explained by identifiable shortcuts in the image model, that is, spurious correlations that inflate apparent performance and can lead to group-dependent errors. Given the observed disparities, we tested whether post hoc per-group decision thresholds can improve overall performance and reduce group-wise gaps. While prior work suggests that post hoc methods can avoid levelling down compared with in-training correction methods, per-group thresholds yielded no significant gains over a single global threshold in our experiments. We resolved this apparent tension through simulation by deriving when per-group thresholding can help, showing that it is effective only when class separation is low and subgroup prevalence differs strongly. These conditions also imply that, when separation is high or prevalence differences are modest, per-group thresholding is unlikely to improve performance, consistent with our empirical results. Finally, we investigated whether adding fairness constraints can mitigate the observed disparities. We found that improvements observed during model development often did not generalise to held-out test data and that outcomes depended strongly on the chosen objective and constraint. Together, these findings provide a framework to identify when targeted interventions, including per-group thresholding and fairness constraints, are likely to add value.

Contents

6.1	Introduction	139
6.2	Related work	140
6.3	Methodology	143
6.4	Results	149
6.5	Discussion	154

6.1 Introduction

This chapter addresses the robustness side of the overarching question in [Chapter 1](#), asking how subgroup disparities can be identified and mitigated once multimodal fusion is in place. It follows directly from the fusion framework of [Chapter 5](#) and from the review ([Chapter 2](#)), which stressed evaluation across patient subgroups and generalisability as conditions any multimodal clinical model must meet.

One obstacle to deploying machine learning-based decision support systems in clinical care is the risk of subgroup performance differences and group-dependent miscalibration, which can increase existing disparities in care ([Alloula et al., 2025](#); [Seyyed-Kalantari et al., 2021](#)). Such issues can be hard to detect from aggregate metrics because a model may achieve strong average performance while concentrating errors in specific patient groups, leading to uneven clinical risk.

A recent trend in machine learning research for healthcare is multimodal modelling, where data from multiple modalities are fused ([Topol, 2023](#)). Multimodal fusion can shift subgroup disparities, so fairness must be evaluated explicitly in multimodal models ([Zhou et al., 2021b](#)). In these settings, group differences may be amplified when the same demographic information is present in more than one modality and is effectively counted twice. We therefore examined whether each modality contributes information beyond what is already available in the other modalities. If tabular attributes can be inferred from images, adding them explicitly may yield limited benefit and instead introduce demographic redundancy.

Because disease prevalence differs strongly between age and sex subgroups and both demographic features are well inferable from chest X-rays in our cohort (see [Table 35](#)), we hypothesised that group-dependent threshold offsets can arise. We treated per-group decision thresholds as both a diagnostic and a potential correction. To clarify why such corrections do not always yield accuracy gains, we used simulation experiments to characterise when this kind of intervention adds value, focusing on model performance

and subgroup prevalence differences.

This work hence had two central research aims:

Aim 1: *Determine whether multimodal late fusion introduces demographic redundancy and contributes to subgroup performance differences.*

Aim 2: *Evaluate whether post hoc calibration and fairness constraints reduce subgroup disparities and identify when such interventions are expected to help.*

The core contributions of this chapter are:

- **Subgroup differences in multimodal ICU prediction.** We quantified discrimination for cardiovascular condition prediction from late fusion and report subgroup performance differences and group-dependent miscalibration that are not visible from aggregate metrics.
- **Per-group thresholding and its expected limits.** We compared post hoc per-group decision thresholds with a single global threshold. As per-group thresholds did not yield measurable gains, simulation results clarified when they can help, mainly when class separation is low and subgroup prevalence differs strongly, which explains the empirical findings and gives a practical rule of thumb.
- **Fairness constraints and generalisation to held-out test data.** We evaluated fairness objectives and constraints with OxonFair (Delaney et al., 2024). We found that apparent improvements during model development often did not carry over to held-out test data and outcomes depended strongly on the chosen objective and constraint.

6.2 Related work

Subgroup disparities. Deploying AI in sensitive healthcare settings requires models that behave fairly across patient groups and that base predictions on clinically relevant information. However, work on algorithmic fairness has shown that strong average performance can mask large subgroup differences. For example, *Gender Shades* (Buolamwini and Gebru, 2018) reported markedly higher error rates for dark-skinned women in commercial facial analysis systems, highlighting intersectional performance gaps. In medical imaging, *Reading Race* (Banerjee et al., 2021) showed that self-identified race

can be inferred from X-ray images, indicating that image models can encode demographic signals even when these attributes are not provided explicitly. Together, these findings motivate routine subgroup evaluation and explicit checks for demographic encoding in clinical AI. Building on this line of work, [Seyyed-Kalantari et al. \(2021\)](#) studied chest X-ray models and reported disparities in missed diagnoses across demographic and socioeconomic groups. They showed that under-served populations, including Black patients and individuals with lower socioeconomic status, experienced higher rates of missed diagnoses and proposed a framework for subgroup-specific evaluation in clinical AI. Similarly, [Alloula et al. \(2024\)](#) analysed a retinal disease classifier trained on UK Biobank data and reported performance differences across demographic groups and assessment centres.

Assessing bias sources. Identifying the source of disparities is important for mitigation, but challenging. [Jones et al. \(2024\)](#) identified three causal sources of dataset bias: *prevalence disparities*, *presentation disparities* and *annotation disparities*. Each tends to require a distinct mitigation strategy; selecting one that does not match the underlying cause can worsen outcomes. Furthermore, proving whether a model truly encodes protected characteristics (like race or gender) to drive prediction is not straightforward.

[Glocker et al. \(2021\)](#) showed that proxy tests based on transfer learning can be insufficient to establish whether such attributes drive predictions and recommended a combined approach of test set resampling, multi-task learning and representation inspection, to assess how protected characteristics are encoded and how this relates to subgroup performance.

[Yang et al. \(2024\)](#) investigated shortcuts as a potential source of unfair outcomes. In shortcut learning, models rely on spurious or easily available cues that correlate with outcomes or protected group membership rather than pathology. They showed that medical imaging classifiers can encode demographic attributes such as race, sex and age, that stronger demographic encoding can be associated with larger fairness gaps and that mitigation methods can reduce in-distribution fairness gaps while maintaining

overall performance, but these gains often do not transfer to external test sets under distribution shift.

Fairness mitigation and calibration. To address fairness concerns, many studies proposed algorithmic constraints during training (Agarwal et al., 2018; Chen et al., 2021a; Goel et al., 2018; Kamishima et al., 2012; Zafar et al., 2017; Zhao et al., 2022; Zhu and Wang, 2022). However, their effectiveness has been questioned. Mosquera et al. (2021) and Pfohl et al. (2022) argued that fairness-aware training does not necessarily improve outcomes and may lower performance. Zietlow et al. (2022) showed that some fairness interventions can reduce accuracy across groups (including protected ones), highlighting the risk of ‘levelling down’. Similarly, Mittelstadt et al. (2023) highlighted that some interventions may unintentionally even harm the groups they aim to support, advocating for ‘levelling-up’ strategies using group-specific threshold calibration tailored to error types. Nonetheless, as Seyyed-Kalantari et al. (2021) note, even threshold-based corrections face challenges: unreliable estimation for small or intersectional groups, increased computational burden when accounting for multiple protected attributes, reliance on socially constructed variables such as race and ethical concerns if performance degrades in specific subgroups. Moreover, such methods may not correct differences in prevalence and can introduce new disparities in other metrics. Still, post-hoc calibration remains attractive because it can be applied to an already trained model and tuned to specific group-wise error trade-offs without retraining the underlying network.

To address some of the limitations, Delaney et al. (2024) introduced *OxonFair*, a toolkit for assessing and mitigating subgroup fairness across tasks in tabular data, computer vision and natural language processing. OxonFair supports a broad set of metrics and constraints and offers a practical, extensible framework for integration into model development pipelines. As a post-processing strategy (vs pre-processing and in-processing), it seeks to avoid ‘levelling-down’ by using per-group threshold calibration towards group fairness (Delaney et al., 2024; Mittelstadt et al., 2023). In this work, we favour validation-based per-group thresholding as a post-hoc method to control

group-wise error trade-offs while keeping the trained model fixed (Mittelstadt et al., 2023). However, it is currently unclear when per-group threshold calibration improves over a single global threshold, particularly in multimodal settings.

Addressed gaps. Existing literature highlights the risks of subgroup differences in medical imaging and the potential of post-hoc calibration, but three key gaps remain for multimodal critical care models. First, the mechanism of redundancy is under-explored: It is often assumed that multimodal models benefit from tabular demographics, but it is unclear if they merely double-count signals already present in the images. Second, the conditions for calibration impact are vaguely defined: Prior work advocates for calibration, but we lack a precise understanding of the conditions (defined by class separation and prevalence) under which per-group thresholds actually improve over global thresholds. Third, the stability of fairness constraints is often assumed from validation results, without rigorously testing whether ‘fair’ validation thresholds satisfy constraints on held-out test sets. We address these gaps by applying the frameworks from Glocker et al. (2021) and Delaney et al. (2024) to check for redundancy, by using simulation to define the exact conditions where calibration adds value and by evaluating whether validation-selected fairness constraints hold on held-out test data.

6.3 Methodology

6.3.1 Data and study population

This study used X-ray images, ECG waveforms and tabular patient information from the MIMIC-IV (version 3.0) database (Goldberger et al., 2024; Johnson et al., 2024, 2023b) following the same data processing as in the previous chapter. This resulted in 110,852 X-ray images (in the following referred to as ‘all data’), 298,270 ECG recordings and 28,262 laboratory tests, vital signs and demographic information (LVD) data that were used to train and evaluate the single modality models, with an overlap of 5,280 patients that were used for the fusion experiments (in the following referred to as ‘fusion data’). The demographics are shown in Table 34. The cohort comprises primarily older,

white patients with a roughly equal distribution between genders. [Table 34](#) also shows the number of positives for the different groups used in the experiments. Using five-fold cross validation, this resulted in partially small positive group sizes, a common problem in fair machine learning that can cause generalisation issues ([Delaney et al., 2024](#)).

6.3.2 Models

This work builds on the models and training pipelines that were developed in the previous chapter. For the X-ray image models, this primarily refers to the ResNet50 architecture, and for ECGs to the spectrogram-based transformation approach ([Krones et al., 2024](#)). For the tabular LVD data, a fully connected neural network was used. For the fusion experiments, we applied the late-fusion approach in which the predicted probabilities of the single-modality models were averaged ([Krones et al., 2025](#)).

6.3.3 Experiments on subgroup differences

Assessing group differences. To test for subgroup differences we used the overall performance drop, worst-group performance and the maximum performance gap ([Alloula et al., 2025](#); [Glocker et al., 2021](#)).

Assessing score offsets. To test whether demographic redundancy is associated with group-dependent score offsets (miscalibration), we used OxonFair ([Delaney et al., 2024](#)) to select per-group decision thresholds on held-out validation data (maximising accuracy without fairness constraints). We compared these to a single global threshold on the test set. If per-group thresholds yielded significant gains, this would indicate that the optimal operating point differs by group, consistent with group-dependent score offsets that can arise from demographic double counting.

Assessing feature encoding. A key challenge in multimodal learning is to determine whether a model relies on spurious or readily available cues as shortcuts. In late fusion, it can be hard to tell whether each modality contributes genuinely new information, or whether some signals are effectively double-counted. If an image model can already infer certain tabular features, such as sex or age, the benefit of explicitly including

those features as tabular inputs becomes questionable. This, in turn, can amplify group-dependent score offsets or miscalibration and concentrate errors in specific patient groups. To determine whether tabular features are implicitly encoded in the image model, we combined the Supervised Prediction Layer Information Test (SPLIT) with multi-task learning as suggested by [Glocker et al. \(2021\)](#). We selected age and sex as the most salient demographic features based on SHAP values from a tabular-only baseline ([Lundberg and Lee, 2017](#)).

6.3.4 Simulation experiments on threshold calibration

To explain why per-group threshold optimisation often failed to improve over global thresholds in our empirical results, we used simulated data to identify the specific conditions where calibration adds value.

Let $G \in \{0, 1\}$ denote a binary group and $Y \in \{0, 1\}$ a disease label. We use $P(\cdot)$ to denote probabilities; for example, $P(G=1) = 1/2$ in our setting. A model produced a continuous disease score $S \in \mathbb{R}$ for each individual. For each simulated individual we drew $G \sim \text{Bernoulli}(1/2)$ and then drew the disease label according to the group-specific prevalence $\pi_g = P(Y=1 \mid G=g)$, with π_0 and π_1 varied on a grid. Conditional on Y , we generated a score from a Normal distribution with fixed separation that did not depend on group:

$$S \mid Y=1 \sim N(\mu_1, 1), \quad S \mid Y=0 \sim N(\mu_0, 1), \quad \mu_1 - \mu_0 = \text{sep}.$$

In all experiments we set $\mu_1 = +\text{sep}/2$, $\mu_0 = -\text{sep}/2$. We transformed scores to two-class probabilities using a group-agnostic logistic (sigmoid) mapping.

We compared two post-hoc policies using the FairPredictor from OxonFair with thresholds learned on validation data only: a global policy in which we assigned all validation samples to a single pseudo-group and learned one threshold t and a per-group policy in which we provided true group labels and learned thresholds t_0 and t_1 . We evaluated both policies on an independent test set. For each prevalence pair (π_0, π_1) we drew

an i.i.d. validation set of size n_{vali} and a test set of size n_{test} , fitted thresholds on the validation set and computed overall test accuracy. Our primary outcome was the accuracy improvement

$$\Delta = \text{accuracy}_{\text{per-group}} - \text{accuracy}_{\text{global}},$$

reported in proportion units and visualised in percentage points (pp) for interpretability.

We explored a regular 70×70 grid with $\pi_0, \pi_1 \in [0.1, 0.8]$. For each grid cell we set a fixed random seed to ensure reproducibility, generated $n_{\text{vali}} = 10,000$ validation and $n_{\text{test}} = 10,000$ test instances and computed Δ . In a second experiment we fixed $\pi_0 = 0.3$, varied π_1 over $[0.1, 0.8]$ and varied the class-separation parameter sep over a grid. This allowed us to visualise how the distribution of validation probabilities p evolved with increasing separability at zero prevalence difference and to quantify how accuracy gains from per-group thresholds depended jointly on prevalence in group 1 and separability.

To validate the observations from the simulated data, we applied ECG-based models for hypertensive diseases using age groups (from the previous chapter) as an example in which the groups exhibit substantial prevalence differences (> 30 percentage points) while the model performance shows weak separability ($< 70\%$ AUC).

6.3.5 Ablation experiments using logistic regression

In late fusion, averaging disease scores implicitly assumes that the contribution of each modality is uniform across demographic groups. While our threshold calibration experiments introduce group-specific decision thresholds, they leave the underlying scores unchanged. This acts only as a shift in the intercept, effectively correcting for differences in group-specific base rates.

To check whether subgroup differences require more than simple threshold shifts, we fitted logistic regression models on top of the modality-specific predictions. While threshold calibration only adjusts the intercept, logistic regression can re-estimate how each modality is weighted per group. We used the predicted probabilities from single

modality models (S) and the binary group indicators (G) as features. The model allows the weighting of the modality to vary by group via the interaction term β_3 :

$$\text{logit}(P(Y = 1|S, G)) = \beta_0 + \beta_1 S + \beta_2 G + \beta_3(S \times G)$$

This provides a formal test: if the logistic regression ensemble does not outperform simple calibrated thresholds, it suggests that subgroup differences are largely explained by base rate shifts (implying $\beta_3 \approx 0$) rather than complex group-dependent miscalibration.

6.3.6 Experiments on fairness

To mitigate observed disparities, we applied OxonFair (Delaney et al., 2024) for post-hoc threshold calibration with fairness constraints. OxonFair was selected because it allows the definition of minimum fairness thresholds and supports per-group thresholding, which has been shown to be optimal under specific assumptions for certain fairness metrics (Delaney et al., 2024). In contrast to calibration-based methods, many mitigation strategies applied during model training are associated with a reduction in overall performance (‘levelling down’) (Mittelstadt et al., 2023) and their effectiveness is highly dependent on the subgroup definition (Alloula et al., 2025). In the absence of clear knowledge about the mechanisms causing bias (cf. Jones et al. (2024)), such methods may not be advisable (Alloula et al., 2025). Therefore, we opted for post-processing correction to improve model fairness while maintaining overall performance (‘levelling up’) (Mittelstadt et al., 2023). OxonFair implements this by searching for a solution along a Pareto frontier that optimises an objective (e.g. accuracy) while satisfying a constraint (e.g. a difference in recall). To construct the frontier, OxonFair performs a computationally optimised grid search over per-group decision thresholds and extracts threshold-based metrics such as TP, FN, FP and TN. As the objectives and fairness constraints of OxonFair are threshold-dependent, we used accuracy as the objective with optional recall difference (equal opportunity) as the constraint, limited to 0.02. Besides optimising accuracy under an equal opportunity constraint, we also optimised the F1-score and enforced a minimum per-group recall (Ryström et al., 2025) to study

clinical ‘levelling-up’ regimes.

6.3.7 Ethical approval

Prior to the start of the project, ethical approval (SSH.OII.CIA.22.108) was obtained through the departmental research ethics committee using the CUREC 1A form, following the University of Oxford’s Central University Research Ethics Committee (CUREC) guidelines. Although the study relied exclusively on secondary data analysis of the publicly available, de-identified MIMIC-IV database and did not require additional ethical approval, this step reflects a commitment to responsible research conduct. Access to MIMIC-IV was granted after completion of the required training on research with human participants and acceptance of PhysioNet’s Data Use Agreement, which mandates safeguarding the data, refraining from any reidentification attempts and adhering to data security standards.

6.3.8 Implementation details

For the neural network models, we used a Binary Cross Entropy loss and the Adam optimiser (Kingma and Ba, 2014). All images were resized to a resolution of 224×224 , using random cropping, rotation and flipping during training. A learning rate of 0.00005 was set and `ReduceLROnPlateau` was used as learning rate scheduler with minimum rate of 0. All models were fine-tuned for 30 epochs. A single NVIDIA H100 80GB HBM3 GPU was used with CUDA version 12.7. If not otherwise stated, the term significant denotes results with a p-value of less than 0.05 with all statistical tests being two-sided; to assess performance differences between models evaluated on identical cross-validation folds, we used paired-samples t-tests on fold-level performance metrics, thereby accounting for the dependence induced by shared data splits.

6.4 Results

6.4.1 Subgroup differences

Figure 19 and Figure 20 show that age and sex are frequently among the most informative predictors across cardiovascular conditions. For chronic ischaemic heart disease, Figure 43 shows in more detail that especially age is a strong discriminator. The mean age in the negative class is 62 years, while in the positive class it is 73 years, a statistically significant difference (two-sample t-test). An age threshold that minimises entropy is 55.5 for the fusion data and 63.5 for the imaging data. Below this threshold, the positive rate is 0.09 for fusion and 0.13 for imaging; above it, the positive rate is 0.35 for both. Regarding sex, the positive rate is 0.32 for males and 0.23 for females in the fusion data, and 0.30 and 0.21 in the imaging data, also a significant difference (chi-squared test). Males constitute 63.8% of the disease-positive group but only 52.7% of the negative group in the fusion data; in the imaging data, the corresponding values are 64.8% and 52.6%.

As shown in Table 16, we also observed significant performance differences between those groups: the model performed better for younger than for older patients and better for males than for females. This is notable because, when compared with Table 34, the two comparisons exhibit opposite distribution patterns: younger patients form the smaller group whereas male patients form the larger group; younger patients have fewer positive observations than older patients, whereas males have more than females. As Medicare primarily covers older patients, its performance mirrors the older group, showing lower performance than other insurance types.

6.4.2 Demographic redundancy

Both demographic features, sex and age, can be predicted with high accuracy from the X-ray images (see Table 35 and Adleberg et al. (2022)). We hence first tested if potential shortcuts learned by the image model could be an issue. However, as shown in Table 35, the multi-task model did not outperform the single-task model for either sex

Table 16: **Fusion models: Performance by subgroup.** Shown is the mean (SD) AUC over different splits for different groups of the original late fusion model. Values are shown for chronic ischaemic heart disease. The results are reported on the held-out test data.

Group	AUC original
Overall	.824 (.016)
Max gap	.111
Worst	.750 (.023)
Age - young	.861 (.034)
Age - old	.750 (.023)
Sex - female	.789 (.010)
Sex - male	.841 (.023)
Insurance - Medicaid	.848 (.053)
Insurance - Medicare	.769 (.024)
Insurance - Other	.846 (.123)
Insurance - Private	.865 (.020)

or age. Additionally, the performance of the SPLIT model did not reach that of the multi-task model. Taken together, these results suggest that disease predictions from images do not rely directly on sex or age information (Glocker et al., 2021).

6.4.3 Threshold calibration and logistic regression

To test whether subgroup differences are associated with group-dependent score offsets, we compared per-group threshold selection with a single global threshold using an accuracy-only objective (no fairness constraint). Compared with the global, accuracy-optimised baseline, these per-group thresholds did not yield a significant accuracy improvement. Table 17 shows this for the example of sex: comparing the ‘Acc optimized’ and ‘Acc optimized Sex’ rows for the ‘Fusion model’ experiments shows that group-dependent thresholds do not significantly improve performance. For the example of chronic ischaemic heart disease, Table 36 confirms these findings for age groups as well, where the global threshold performance of 0.788 only improved slightly to 0.791 (compare the first column ‘Overall’ result for experiment b) with the first column ‘Sex’ result for experiment c)). To understand why this occurs, we simulated a range of prevalence levels and model performance scenarios.

Table 34 shows that the prevalence difference between age groups was 22 percentage

Table 17: **Threshold calibration.** Shown is the mean (SD) over five folds for different metrics and diseases for different threshold optimisation methods. Results are reported on the held-out test data. The ‘Affected samples’ rows show the ratio of affected samples by having group specific thresholds. The absolute prevalence difference between the groups is reported on the total dataset which was split randomly into train, validation and test. For logistic regression, the scores from the single modalities were used as features in a logistic regression model together with the corresponding group identifiers. For ‘Acc log. reg.’ the final model probabilities were combined with the one specific group identifier; for ‘Acc log. reg. all’ the single modality probabilities were used; for ‘Acc log. reg. mul.’ the group identifiers for sex, age and race were used. 1 = Hypertensive disease, 2 = Ischaemic heart disease, 3 = Chronic ischaemic heart disease, 4 = Cardiomyopathies disease, 5 = Dysrhythmias disease, 6 = Heart failure.

Measure	Group	1	2	3	4	5	6
Fusion model (fusion data)							
AUC	-	.799 (.009)	.799 (.020)	.824 (.016)	.809 (.032)	.845 (.013)	.860 (.007)
Acc @ 0.5	-	.724 (.010)	.785 (.022)	.744 (.029)	.933 (.006)	.776 (.016)	.786 (.013)
Acc optimized	-	.758 (.005)	.839 (.012)	.788 (.012)	.950 (.005)	.779 (.020)	.804 (.007)
Acc optimized	Sex	.761 (.007)	.837 (.011)	.791 (.011)	.953 (.005)	.778 (.021)	.803 (.009)
Affected samples	Sex	.010 (.015)	.011 (.015)	.015 (.016)	.012 (.015)	.013 (.015)	.011 (.014)
Prevalence diff.	Sex	.022	.046	.090	.018	.004	.018
Acc log. reg.	Sex	.759 (.004)	.837 (.012)	.789 (.013)	.953 (.005)	.780 (.019)	.803 (.008)
Acc log. reg. all.	Sex	.758 (.006)	.841 (.012)	.792 (.010)	.950 (.004)	.778 (.016)	.812 (.013)
Acc log. reg. mul.	-	.756 (.009)	.834 (.013)	.784 (.013)	.951 (.007)	.779 (.025)	.800 (.009)
Image model (all image data)							
AUC	-	.741 (.007)	.691 (.019)	.794 (.006)	.721 (.027)	.792 (.012)	.815 (.012)
Acc @ 0.5	-	.675 (.029)	.749 (.030)	.745 (.039)	.928 (.012)	.731 (.012)	.739 (.017)
Acc optimized	-	.715 (.008)	.889 (.006)	.800 (.007)	.944 (.006)	.737 (.010)	.777 (.008)
Acc optimized	Sex	.716 (.008)	.889 (.006)	.800 (.007)	.944 (.006)	.735 (.007)	.778 (.007)
Affected samples	Sex	.015 (.012)	.008 (.011)	.007 (.010)	.006 (.009)	.009 (.013)	.010 (.012)
Prevalence diff.	Sex	.001	.030	.095	.025	.047	.001
Acc log. reg.	Sex	.715 (.008)	.888 (.006)	.800 (.006)	.945 (.006)	.738 (.010)	.777 (.007)
Acc log. reg. mul.	-	.721 (.007)	.889 (.006)	.800 (.006)	.945 (.005)	.741 (.010)	.777 (.007)
ECG model (fusion data)							
AUC	-	.657 (.021)	-	-	-	-	-
Acc @ 0.5	-	.651 (.015)	-	-	-	-	-
Acc optimized	-	.670 (.008)	-	-	-	-	-
Acc optimized	Age	.705 (.009)	-	-	-	-	-
Affected samples	Age	.091 (.067)	-	-	-	-	-
Prevalence diff.	Age	.300	-	-	-	-	-
Acc worst group opt.	-	.573 (.016)	-	-	-	-	-
Acc worst group opt.	Age	.613 (.022)	-	-	-	-	-
Acc log. reg.	Age	.706 (.008)	-	-	-	-	-
Acc log. reg. mul.	-	.702 (.008)	-	-	-	-	-

points; between sex groups it was 9 p.p. The classifiers’ discriminations were good, with AUCs above 0.70 across groups (Table 17). However, in simulations (Figure 23) we found that per-group threshold optimisation yields meaningful gains only when class separation is low and prevalence gaps are large. For model performances and prevalence differences as we observed them in our image and fusion models, the expected accuracy improvements over a single global threshold are small. Using the ECG-based model for

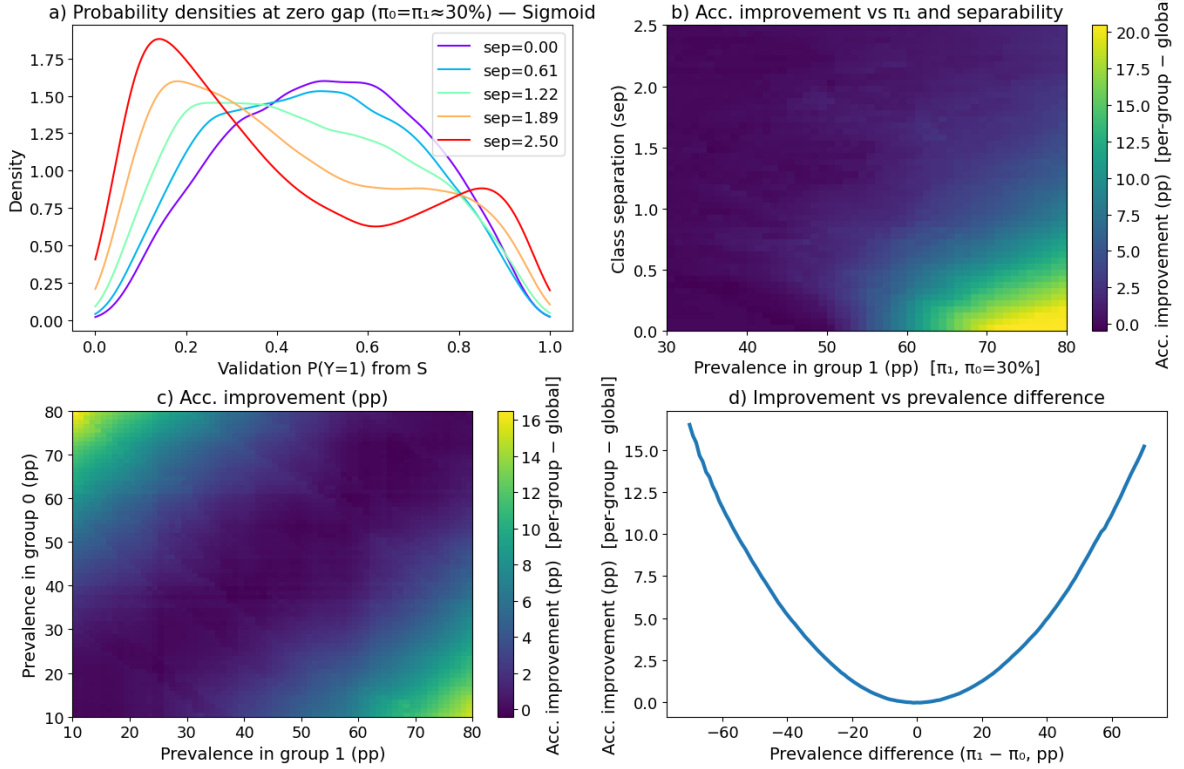


Figure 23: **Simulation: Per-group thresholds vs a single, global threshold.** (a) Validation probability densities for multiple class-separation settings at zero prevalence gap. (b) Accuracy improvement (pp) of per-group over global thresholding as a function of class separation and group-1 prevalence (π_0 fixed at 30%); gains are largest when separation is low and prevalence differences are high. (c) Accuracy improvement (pp) across the prevalence grid (π_0 vs π_1). (d) Accuracy improvement (pp) against prevalence difference ($\pi_1 - \pi_0$); only large prevalence differences yield meaningful gains.

hypertensive diseases, which showed lower class separation than the image-based models examined above, and focusing on age groups which have large prevalence differences, we confirmed the findings from simulations and showed that in this setting group-specific thresholds can yield a significant improvement in model performance (Table 17).

Finally, we assessed whether subgroup differences required more complex corrections than simple threshold shifts. We fitted logistic regression models on top of the modality predictions, allowing both the intercept and the modality weights to vary by group (Table 17). For chronic ischaemic heart disease, the logistic regression ensemble (‘Acc log. reg.’) yielded an accuracy of 0.789, which did not differ significantly from the simple global threshold (0.788) or the group-specific threshold (0.791). This pattern held across all conditions, suggesting that observed disparities are largely explained by

Table 18: **Fusion models: Fairness constraints.** Shown is the mean (SD) of accuracy, FNR and F1 for different fairness interventions on the fusion model, using OxonFair with fairness constraints. Values are shown for chronic ischaemic heart disease. ‘Max gap’ reports the maximum performance gap between either age, sex or insurance subgroups; ‘worst group’ reports the performance of the worst subgroup across them. Details can be found in Table 36.

Objective	Fairness constraint	Metric	Accuracy	FNR	F1
Accuracy	None	Max gap	.150	.248	.110
		Worst group	.714 (.017)	.718 (.116)	.628 (.053)
Accuracy	Equal opportunity of 0.02 (recall or FNR gap)	Max gap	.142	.342	.090
		Worst group	.706 (.028)	.648 (.193)	.612 (.110)
F1	Minimum recall / maximum FNR of 0.5	Max gap	.129	.269	.111
		Worst group	.713 (.017)	.555 (.055)	.634 (.151)

base-rate differences captured by group-specific intercepts, with limited evidence that group-specific modality weights are needed.

6.4.4 Fairness constraints

When per-group threshold calibration was applied without any fairness constraints, overall performance did not increase compared with a global threshold optimisation, as discussed above (see Table 17). When additional fairness constraints (equal opportunity, implemented as an absolute recall-difference bound) were applied while optimising accuracy (see the second row in Table 18), the maximum accuracy gap decreased. However, although the recall-difference bound was enforced on validation data with a maximum difference of 0.02, this threshold did not hold on the held-out test set (where, equivalently, the maximum FNR gap was 0.342). Finally, when optimising F1 instead of accuracy and enforcing a minimum per-group recall rather than recall differences (see the last row in Table 18), the constraints were closer to being satisfied on the test set: the worst-group FNR was now only slightly above the 0.5 threshold, at 0.555.

6.5 Discussion

6.5.1 Main findings

This study set out to characterise subgroup differences and evaluate mitigation strategies. Five findings emerged. First, significant subgroup differences exist, particularly across age and sex. Second, per-group threshold optimisation did not improve overall performance compared with a single, globally optimised threshold. However, rather than simply attributing this to good initial calibration, our simulation and logistic regression experiments provide a mechanism-based explanation: under the observed conditions of relatively high class separation and moderate prevalence differences, group-specific thresholds are unlikely to yield accuracy gains. As we showed, such interventions effectively add value only when separation is low and prevalence differences are high. Third, under fairness-constrained calibration selected on validation data, the test set recall gap exceeded the level enforced on validation data. When optimising the F1 score instead of accuracy and enforcing minimum recall rather than a recall-gap bound, the constraints were closer to being satisfied on the test data. Fourth, when models trained up to a historical cut-off were evaluated on later years, we did not observe a steady temporal decline; instead, performance varied year by year, which supports ongoing monitoring rather than assuming monotonic decay (see [Figure 45](#)). Fifth, while age and sex are informative and both are predictable from X-ray images, comparisons of single-task, multi-task and SPLIT variants showed no evidence that disease prediction relies on these attributes (see [Table 35](#)).

6.5.2 Limitations

The considered subgroups focused on age, sex and insurance, while intersectional and site-level effects remain under-explored; expanding protected attributes and pooling multi-site data are practical next steps that do not alter the main conclusions. OxonFair and related calibration methods depend on reliable group labels and can be unstable for small groups; these issues may be mitigated by larger samples or by sharing thresholds

across similar groups. Simulations relied on Gaussian score assumptions which, while illustrative, may not fully capture the complexity of clinical score distributions. Temporal analyses relied on anchor-year mappings, which limit fine-grained temporal analysis; this choice protects privacy and could be revisited with richer timestamps in future datasets. Finally, results reflect internal validation; external, multi-site prospective studies are needed to assess performance and practical deployment.

A more conceptual limitation concerns the question of which groups fairness should be defined over in the first place. Our analysis fixed the groups in advance, using socially salient attributes such as age, sex and insurance type. This implicitly equates the groups that matter for fairness with groups that are socially recognised, yet a second, error-based notion is equally defensible, namely the groups that the model actually misclassifies, which need not align with any predefined demographic category (Alloula et al., 2025). The two views overlap only partially. Where a socially vulnerable group is also the one most often misclassified, the choice of framing does not matter; where they diverge, optimising fairness for predefined attributes can leave a harmed but unlabelled subpopulation unprotected, while chasing every error-defined cluster risks fitting noise and offers no social or legal grounding. A further difficulty is that the mechanism behind a disparity is often unobserved: if a systematic bias acts through a variable we do not record, or through an unrecognised proxy, both subgroup analysis and post-hoc calibration may leave it undetected, because we can only correct for the disparities and attributes we choose to measure (Jones et al., 2024). We do not resolve this tension here, but flag it as a reason to treat any single subgroup definition as provisional, to combine predefined attributes with data-driven error auditing and, where possible, to reason about the causal pathway generating a disparity rather than only its surface statistics.

6.5.3 Conclusion

Overall, the evidence supports a simple playbook for reliable multimodal prediction in critical care: check for redundancy, identify when post-hoc threshold calibration is expected to help (based on separation and prevalence) and evaluate subgroup trade-offs

under threshold-based and constraint-based mitigations that aim for ‘levelling up’. Future work should widen subgroup coverage, study intersectional fairness, link thresholds to clinical use and safety preferences and validate externally with ongoing monitoring. A specific priority is to study extreme cases, that is, settings at the boundaries of the conditions identified here. Our simulations located the regime in which per-group calibration adds value, namely low class separation combined with large prevalence differences, yet our cohort sat in a milder regime. Future work should therefore target subgroups and tasks with extreme prevalence imbalance, very low separability or very small group sizes, where group-dependent errors and the gains from targeted calibration are expected to be largest and where the stability of fairness constraints is most likely to be stressed.

7 Discussion

Contents

7.1	Principal findings	157
7.2	Strengths and limitations in context	159
7.3	Implications	163
7.4	Future research	165

7.1 Principal findings

The central question this thesis set out to answer is: *Under which conditions do multimodal approaches improve the prediction of cardiovascular conditions in ICU patients?*

The specific conditions are discussed below, but the short answer is that multimodal fusion improves predictive accuracy only conditionally and that understanding those conditions is more clinically valuable than the performance gains themselves. Three findings, each arising from a different stage of the work, converge on this conclusion.

The first is *architectural*. This thesis compared early, intermediate and late fusion techniques across six cardiovascular conditions and four data modalities, the latter made possible by the digitisation pipeline and generalisable waveform modelling approach. Despite the theoretical appeal of more complex strategies that model cross-modal interactions, late fusion, the simplest of the three, proved most robust across conditions and data splits. This might be viewed as counterintuitive: in a domain where richer models are generally expected to outperform simpler ones, the advantage went to the approach with fewer assumptions about how modalities relate to each other.

The second is *informational*. Uncertainty-based analysis showed that the benefit of adding modalities was concentrated in cases where single-modality predictions were least certain. Temporal analysis further revealed that the first 24 hours of ICU admission represent the window during which accumulating data most improved predictions, largely

because laboratory results typically became available during this period. Taken together, these findings reframe multimodal fusion as a targeted tool rather than a blanket improvement, where its value depends on what the existing data already say and when they say it.

The third concerns *equity*. Per-group threshold optimisation did not yield improvements in the empirical setting, and simulation experiments showed that such calibration only helps under specific conditions, namely low class separation combined with large prevalence differences. Fairness constraints that appeared effective on validation data did not reliably transfer to test data.

These three findings map back onto the two subsidiary questions posed in [Chapter 1](#). The architectural and informational findings together answer SQ 1, on how fusion strategies compare and under which conditions multimodal integration outperforms single-modality baselines: late fusion is the most robust strategy under ICU data constraints, and its benefit is concentrated in uncertain cases and within the first 24 hours of admission. The equity finding answers SQ 2, on how subgroup disparities can be identified and mitigated: redundancy and miscalibration can be diagnosed, but per-group calibration and fairness constraints help only under specific conditions and do not always transfer to test data. The digitisation and waveform-integration work in [Chapter 3](#) and [Chapter 4](#) sits one level below these subsidiary questions: rather than answering SQ 1 or SQ 2 directly, it makes the overarching question askable at all, by turning paper-bound and expert-dependent waveforms into machine-readable inputs that the fusion experiments require. The digitisation of paper ECGs in particular is therefore not a stand-alone tool but the first link in answering the overarching question, since without scalable waveform representations the multimodal integration studied in [Chapter 5](#) and [Chapter 6](#) would be restricted to the modalities that are already digital. Both subsidiary answers, together with this representational foundation, feed back into the overarching question, by making explicit the architectural, informational and equity conditions under which multimodal integration can be trusted.

What unites these findings is a shift in the central question, from whether multimodal integration helps to when, how and for whom it does so. Answering these questions, rather than simply optimising aggregate performance, is what a principled framework for multimodal clinical prediction should be designed to do.

7.2 Strengths and limitations in context

7.2.1 From individual stages to an end-to-end pipeline

Most prior work on multimodal clinical applications addressed only one stage of the data pipeline in isolation. Digitisation studies focused on signal reconstruction without downstream modelling (Fortune et al., 2021). Fusion studies combined a small number of modalities for a single condition (Huang et al., 2020c). Fairness analyses were conducted mostly for single-modality models (Seyyed-Kalantari et al., 2021). Existing processing pipelines for the MIMIC database reflected this fragmentation. HAIM-MIMIC-MM (Soenksen et al., 2022) integrated imaging and tabular data but excluded ECG waveforms entirely and relied on software requirements that limited reproducibility. MIMIC-IV-Ext-MDS-ED (Lopez Alcaraz and Strodthoff, 2024) included ECGs and tabular data but was restricted to the first 1.5 hours after emergency department arrival and excluded imaging data such as chest X-rays. Symile-MIMIC (Saporta et al., 2025) incorporated X-rays, ECGs and laboratory tests but was not specific to the ICU and required all three modalities to be present, restricting its cohort.

Addressing the full pipeline in a single body of work provides a more complete picture of the challenges involved in building multimodal clinical prediction systems. The digitisation and waveform integration work in Chapter 3 and Chapter 4 was motivated by the observation that reliable waveform representations were a prerequisite for meaningful fusion experiments, a dependency that studies beginning at the fusion stage typically take as given. Building on the spectrogram-based waveform approach developed in Chapter 4, the pipeline in Chapter 5 combines ECG waveforms with chest X-rays, clinical notes and structured tabular records, aligned temporally in an event-based format. It

is designed to be reproducible and to be published as a PhysioNet Software Project following peer review.

However, covering the full pipeline in one body of work comes with some limitations. The core fusion and fairness experiments were conducted retrospectively on a single data source, MIMIC-IV (Johnson et al., 2023b), collected at one teaching hospital in Boston, using ICD-based diagnostic labels that are assigned for billing rather than clinical research (Dragan et al., 2022), which introduces a potential source of label noise. The individual pipeline components were tested more broadly, with the digitisation and spectrogram-based waveform models evaluated on external datasets in Chapter 3 and Chapter 4, though both exhibited a gap between cross-validation scores and performance on hidden test sets. Without clinical validation or external replication of the pipeline, demonstrating that it can be assembled end to end is not the same as showing that it generalises across sites. That said, the modular design allows individual components to be retrained on more diverse data without redesigning the overall approach.

7.2.2 What fusion strategies can and cannot do

Earlier comparisons of early, intermediate and late fusion have consistently been restricted to one clinical task or a narrow set of modalities. Huang et al. (2020c) compared fusion strategies for pulmonary embolism detection using clinical data and CT scans; Imrie et al. (2025) focused on dermatological classification from images and tabular data; Holste et al. (2021) examined breast cancer classification with clinical data and MRI and Hayat et al. (2022) evaluated mortality and phenotyping using clinical data and X-rays. Since the relative advantage of a fusion strategy might vary by condition, each comparison is task-specific, limiting broader conclusions about strategy selection.

In Chapter 5 we tested all three strategies across six cardiovascular conditions at once using four different modalities. Late fusion produced consistently better results than other fusion techniques. The breadth matters because the relative advantage of a fusion strategy can depend on the condition being predicted, a nuance that single-task studies cannot capture. That late fusion performed most robustly across all six conditions lends

greater confidence to this finding. This is broadly consistent with earlier medical studies that also favoured late fusion (Huang et al., 2020c), but it contrasts with studies that reported advantages for early or intermediate fusion in settings with larger or more homogeneous datasets (Guo et al., 2019). The point is not that those studies were wrong. With more data, lower missingness and more consistent alignment across modalities, richer fusion strategies may genuinely perform better. The heterogeneity and missing data typical of ICU settings represent a specific constraint under which the benefits of joint optimisation are outweighed by its sensitivity to incomplete inputs (see Table 10). This shifts the interpretation away from late fusion being universally preferable, toward late fusion being preferable in this particular data environment.

Moreover, the uncertainty analysis in Chapter 5 goes beyond overall performance metrics by showing where fusion helps most. We ranked cases by single-modality certainty and quantified the gain from adding further modalities, following the performance-gain approach of Imrie et al. (2025). Most fusion studies report average performance across all cases, which may suggest that gains are spread evenly. But the benefit of multimodal data appears to be concentrated in cases where the model is most uncertain. This cannot be seen from aggregate comparisons alone. It also changes the practical question from whether to collect more data for a given patient, to when doing so is most likely to matter. The temporal analysis adds another layer to this, finding that the first 24 hours of ICU admission is the window where accumulating data improves predictions most, consistent with previous findings on the timing of data availability in critical care (Deasy et al., 2020; Thiele et al., 2025). That is a level of operational detail that aggregate comparisons cannot provide (Sivaraman et al., 2023).

But neither analysis is without limits. The uncertainty estimates come from the models themselves and have not been tested against dedicated calibration methods and the temporal finding reflects when laboratory results happen to become available rather than something fundamental about the fusion strategy itself, meaning it may not hold where turnaround times differ. The modality coverage, though wider than in most prior work, remains incomplete: computed tomography, wearable sensor data and echocardiography

were excluded due to limited availability in MIMIC-IV at the time of the experiments. The recent release of the MIMIC-IV-ECHO module (Gow et al., 2026), which links structured echocardiographic measurements to the MIMIC-IV database, now makes it feasible to address this gap. More broadly, all fusion experiments were conducted retrospectively on data from a single institution. Although the efficiency curves in Chapter 5 offer a principled way to think about selective data collection, turning them into a clinical decision support system that requests additional tests in real time remains a goal for future work. Ideally multi-site validation would be needed to confirm whether the observed patterns hold in practice.

7.2.3 From observing disparities to diagnosing them

Most prior fairness work in clinical machine learning has focused on documenting subgroup performance differences in single-modality settings. Seyyed-Kalantari et al. (2021) showed that chest X-ray models exhibited higher rates of missed diagnoses for underserved populations. Banerjee et al. (2021) demonstrated that demographic attributes such as race can be inferred from medical images and Alloula et al. (2024) reported performance differences across demographic groups in a retinal disease classifier. These studies established that subgroup disparities exist, but they were typically conducted on single-modality systems and did not examine when corrective interventions are expected to work. On the mitigation side, several studies proposed in-training fairness constraints (Zafar et al., 2017), but these have been associated with levelling-down effects that can reduce accuracy across groups, including the groups they aim to protect (Mittelstadt et al., 2023; Zietlow et al., 2022). Yang et al. (2024) showed that fairness improvements achieved during model development often did not transfer to external test sets under distribution shift.

The analysis in Chapter 6 extends this body of work in three directions. Using the SPLIT framework of Glocker et al. (2021), it tested for demographic redundancy in a multimodal setting and found that although age and sex could be predicted from chest X-rays with high accuracy, disease predictions did not rely directly on these features.

Simulation experiments then identified that per-group threshold calibration improves over a single global threshold only under low class separation combined with large prevalence differences, conditions not met in our cohort, which provides a diagnostic criterion researchers can apply before committing to post-hoc corrections. Finally, fairness constraints enforced on validation data using OxonFair (Delaney et al., 2024) did not reliably transfer to held-out test data, though optimising the F1 score with a minimum per-group recall constraint (Rystrøm et al., 2025) narrowed this gap, offering practical guidance absent from studies evaluating fairness interventions on validation data alone (Yang et al., 2024).

Nevertheless, these contributions should be read alongside their limitations. The subgroup analysis covered age, sex and insurance status but did not examine intersectional effects or site-level variation and small positive group sizes in some cross-validation folds may have reduced the reliability of group-specific estimates, a common challenge in fair machine learning (Delaney et al., 2024). The simulation experiments, while useful for deriving boundary conditions, relied on Gaussian score distribution assumptions that may not fully capture the complexity of clinical prediction scores. The thesis does not offer a fully satisfactory solution to stable subgroup fairness, but the simulation framework provides a diagnostic tool that researchers can apply before committing to a mitigation strategy.

Understanding *why* the observed patterns emerged is what determines how far they can be expected to generalise and what they imply for practice. This is discussed in the next section.

7.3 Implications

The *dominance of late fusion* is best explained by the data environment in which it operates. ICU data are characteristically incomplete: modalities arrive at different times, recording quality varies between patients and missing values are common. Early and intermediate fusion strategies, which attempt to learn cross-modal representations

jointly, are more sensitive to these irregularities because they often require aligned and complete inputs during training (Huang et al., 2020b). Late fusion sidesteps this problem by training each modality independently and combining predictions only at the final stage, which makes it less expressive in theory but more robust in practice. For system designers working in similar data environments, this suggests starting with the simplest fusion architecture rather than investing in more complex joint-training approaches from the outset. Similar logic applies to the uncertainty findings.

The concentration of multimodal *benefit in high-uncertainty cases* has a similarly intuitive explanation. When a single modality already produces a confident prediction, additional data are largely redundant. It is precisely when the available data are ambiguous, for instance when a chest X-ray is unclear or vital signs fall within a grey area, that a complementary modality can resolve the uncertainty. For clinicians, this suggests that the decision to order additional tests could be guided by model confidence rather than by protocol alone (Tran et al., 2022). A model that flags its own uncertainty could prompt targeted data collection, for example requesting laboratory results earlier in the admission window for patients where imaging-based predictions are uncertain. For system designers, this points toward active decision frameworks in which data acquisition is selective rather than exhaustive, reducing cost and patient burden without sacrificing predictive quality.

The *temporal finding*, that the first 24 hours are the most informative window, follows largely from the availability of laboratory results during this period. This has a practical implication for clinical workflows: if laboratory data are the primary driver of improvement within this window, then prioritising earlier collection of laboratory tests at admission could bring forward the point at which reliable multimodal predictions become possible. This is not a recommendation to collect more data indiscriminately, but rather to collect the right data sooner.

The *fairness results* carry implications for both model developers and policymakers. The finding that per-group calibration does not help when class separation is already high and

prevalence differences are modest is not merely a negative result. It provides a diagnostic rule: before committing resources to post-hoc corrections, researchers can assess whether the conditions for their effectiveness are met in their particular setting. The instability of fairness constraints across data splits raises a broader concern. Post-hoc corrections that appear effective on validation data may not hold in deployment (Yang et al., 2024), which suggests that their stability should be tested across multiple data partitions before adoption.

It is worth emphasising what this evidence does not show. Most experiments were conducted on a single dataset from a single institution. The conditions identified here, such as the dominance of late fusion and the limited benefit of per-group calibration, may not hold in different clinical environments or with different patient populations. The findings provide a principled framework for evaluating multimodal systems, but they should not be taken as universal prescriptions.

7.4 Future research

Each of the three main findings, the dominance of late fusion, the concentration of multimodal benefit in uncertain cases and the instability of fairness interventions, raises a question that this thesis could not answer but that follows directly from the evidence.

Regarding the *boundary conditions of fusion architectures*, late fusion outperformed early and intermediate strategies across all conditions tested, but it remains unclear whether this reflects insufficient data volume, representations from modality-specific encoders that are not amenable to joint optimisation, or both. Disentangling these factors, for instance by scaling dataset size while holding architecture constant, or by introducing alignment objectives such as contrastive pre-training across modalities, would clarify when richer fusion strategies become worthwhile. Extending the pipeline to additional modalities, particularly echocardiography now that the MIMIC-IV-ECHO module is available (Gow et al., 2026), would test whether the observed pattern holds for richer modality sets.

The finding that multimodal benefit concentrates in high-uncertainty cases opens the door to *uncertainty-guided data acquisition*, an active decision framework in which a model first assesses its own confidence and requests additional tests when uncertainty exceeds a threshold. Realising such a system would require reliable uncertainty estimates, which in turn calls for a systematic comparison of calibration methods in the multimodal ICU setting. As discussed in [Chapter 5](#), the model could also be extended to simulate positive and negative inputs for a candidate test, evaluating the expected reduction in uncertainty before the test is ordered. Building and validating this type of closed-loop decision support, ideally in collaboration with clinicians, represents a direct extension of this work. Relatedly, the temporal analysis captured only aggregate values within each time window, whereas clinicians often prioritise trends over absolute values. Explicit modelling of temporal trajectories could improve early prediction, though this requires a sufficiently dense recording frequency that is not uniformly met across all modalities and clinical sites.

The instability of validation-selected fairness constraints on held-out test data raises the question of *fairness under distribution shift*, perhaps the most pressing open problem because it undermines confidence in any post-hoc correction deployed in practice. Future work should test whether the simulation-derived boundary conditions for per-group calibration, namely low class separation and large prevalence differences, arise in other clinical populations or prediction tasks, particularly those with greater demographic imbalance than the single-site MIMIC cohort studied here. Extending the subgroup analysis to intersectional groups and to multi-site data, where site-level variation introduces additional confounding, is a further gap. External validation on datasets from hospitals with different patient populations and recording practices is needed more broadly to assess whether the fusion strategies and fairness findings reported here generalise beyond a single institution.

A further direction concerns foundation models and LLM-based multimodal systems, and the more fundamental question of whether task-specific models such as those built here are still needed at all. This thesis trained dedicated models on a single

cohort, yet a growing body of work develops large, pre-trained foundation models for medicine (Bommasani et al., 2021; Moor et al., 2023), including transformer models that predict ICU outcomes directly from longitudinal health records (Renc et al., 2024) and multimodal architectures that combine images, text and structured data (Acosta et al., 2022; Wang et al., 2022a). Such models could act as stronger modality encoders or as a shared backbone for fusion, and LLMs in particular open a route to integrating unstructured clinical notes alongside imaging and waveforms. Recent evidence suggests the gap is narrowing quickly, as an independent evaluation found that general-purpose large language models outperformed specialised clinical AI tools on medical knowledge and real-world clinical queries (Vishwanath et al., 2026). If generalist models continue to match or exceed task-specific pipelines, the choice of which to deploy may rest less on accuracy and more on cost, governance, data privacy and dependence on external, internet-based services, factors that often favour smaller local models in resource-limited or strictly regulated settings. The opposite case, that dedicated models remain worthwhile where data are scarce or highly domain-specific, or are best paired with a generalist in a collaborative setup, is equally plausible (He et al., 2026). Which regime applies, and whether the conditions identified in this thesis still hold once modalities are encoded by a shared foundation model, is an open question that the systematic evaluation framework developed here is well placed to test.

Taken together, these questions share a common thread, each asks not whether multimodal fusion works in principle, but under what conditions it can be trusted in practice. This thesis has shown that answering this question requires looking beyond aggregate performance to the specific architectural, informational and equity conditions that determine whether additional data genuinely help. Addressing these questions is necessary but not sufficient; the path from research to clinical deployment involves further administrative, regulatory and organisational challenges, as well as the integration of such systems into clinical workflows (Rajpurkar et al., 2022; Widner et al., 2023; Yildirim et al., 2024). The framework and evidence presented here provide a starting point for that broader effort.

Financial declaration

In light of the substantial tuition fees required by the University of Oxford and the associated costs of living, this doctoral research was supported by several external sources. These funding bodies had no influence on the conduct or outcomes of the research. I am grateful to McKinsey & Company for the loan provided during the first year of my studies. During the second and third year, my living expenses were partially supported by the German taxpayers through the Friedrich Naumann Foundation, while the Madam Shirley Scholarship partially offset my tuition fees. All remaining financial requirements during those years and all living costs and tuition fees for the years following the third year were self-funded.

A Extended literature

This chapter outlines key technical concepts ([Section A.1](#)), as well as socio-technical aspects ([Section A.2](#) and [Section A.3](#)), which are important for comprehending and effectively applying the findings of this thesis.

Contents

A.1 Technical concepts	169
A.2 AI deployment challenges	180
A.3 Reliability in medical computer vision	182

A.1 Technical concepts

This section provides an introduction to key technical concepts used in this research. To apply imaging models in multimodal applications, they must first be initiated with appropriate starting weights. To determine the most effective approach for pre-training medical imaging models to enhance downstream performance and minimise the need for labelled data, the following two subsections, [Section A.1.1](#) and [Section A.1.2](#), provide an overview of different methodologies. [Section A.1.3](#) and [Section A.1.4](#) explain the two main model architectures used for medical images, CNNs and transformers.

A.1.1 Transfer learning

Transfer learning enables models to leverage knowledge acquired from one task to improve performance on a related task. Transfer learning offers substantial benefits in healthcare AI applications by improving model performance, reducing training time and addressing data scarcity. It allows models to adapt to new tasks, handle multi-modal data and promote interdisciplinary knowledge transfer ([Walker et al., 2022](#)).

In transfer learning, a model is trained on one dataset and later fine-tuned on target data. For example, a network might first be trained on natural images before being fine-tuned on and for specific medical images ([Walker et al., 2022](#)). This can be done by

either fine-tuning the weights of all layers or only specific layers (often the last layer only).

Transfer learning can be divided into two categories: homogeneous and heterogeneous learning (Zhuang et al., 2020). In homogeneous transfer learning, it is assumed that the domains are of the same feature space. Hence, the model only needs to be adapted for distribution shifts. For example, a homogeneous task can be adapting a model trained on high-quality dog images to recognize dogs in low-resolution or poorly-lit images, where the challenge lies in addressing the shift in image quality within the same feature space. In heterogeneous transfer learning, it is assumed that the feature spaces are different; hence, the model requires feature space adaption. For instance, a heterogeneous task could involve adapting a model trained for categorizing text-based news articles into different topics to perform similar categorization on images. This requires the model to adapt across fundamentally different types of data, transitioning from text to image processing.

The choice of the correct adaption or fine-tuning method also comes with the question of computational requirements. Fine-tuning all model parameters for many tasks will demand large memory and computational time. Hence, freezing most of the parameters and only learning a small number of parameters or adding new modules and training them can help to reduce storage and run-time needs (Bommasani et al., 2021). However, how to best choose the parameters and layers to adapt is an open research question. This also holds for the more basic question of how lightweight adaptations of only some layers perform when model size increases (Bommasani et al., 2021). Previous literature has also shown that the fine-tuning strategies are architecture, data and pre-training dependent (Glocker et al., 2022; Khan and Fang, 2024) and choosing the wrong one can lead to a so-called knowledge-forgotten-issue (Gu et al., 2024). For example, if the target domain is significantly different from the pre-training domain, training models from scratch can work better than fine-tuning large pre-trained models (Bommasani et al., 2021). One example is in the legal domain with LegalBERT (Chalkidis et al., 2020) vs Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018).

The concrete effects and how much fine-tuning is necessary is an open research question.

For language-based settings, [Bommasani et al. \(2021\)](#) summarise the advantages of prompting¹³ the downstream domain such as that a good prompt can have the same effect as 100 times more training data and hence increase data-efficiency ([Toutanova et al., 2021](#)). [Nori et al. \(2023\)](#), for example, showed that through advanced prompting, GPT-4 can outperform more specialised models, such as Med-PaLM 2, on specialised tasks such as medical question answering. How prompting can best be applied to other modalities is active research.

While we mostly understand task specialisation under model adaption, the described techniques can also be used to adjust for temporal shifts to keep models up to date ([Bommasani et al., 2021](#)).

A.1.2 Learning strategies

In machine learning, various learning strategies are used to train models, particularly in the field of computer vision. The main learning strategies include supervised, unsupervised, semi-supervised and self-supervised learning ([Prince, 2023](#)):

Supervised learning: In supervised learning, models are trained using a labeled dataset, where each data point is associated with a corresponding label or ground truth. The model learns to map input data (e.g. images) to output labels (e.g. object categories) by minimising the difference between its predictions and the ground truth. Supervised learning is often used in computer vision, for tasks such as image classification and object detection or segmentation.

Unsupervised learning: In unsupervised learning, models are trained on an unlabeled dataset and the goal is to discover underlying patterns or structures within the data. The learning process is driven by the data itself, without any explicit supervision.

¹³Prompting refers to the process of providing a model with a specific input or instruction, which guides the model to generate a desired output or perform a particular task. This technique is crucial in models like GPT, where the prompt determines the context and nature of the model’s response. For example, in a healthcare context, if a clinician inputs ‘List potential side effects of medication X’, this input serves as the prompt. The model processes this prompt to provide a detailed list of possible side effects associated with medication X.

Unsupervised learning is often used in computer vision for tasks such as clustering (grouping similar images together), dimensionality reduction (e.g. using PCA or t-SNE) and generative modeling (e.g. generating new images with GANs).

Semi-supervised learning: Semi-supervised learning is a mix of supervised and unsupervised learning. It usually uses a small amount of labeled data and a much larger amount of unlabeled data to improve the model’s performance compared to using either or. In computer vision, semi-supervised learning can be applied to similar tasks as supervised learning, such as image classification or object detection, to make better use of available data and achieve better performance than with purely supervised or unsupervised techniques alone.

Self-supervised learning: Self-supervised learning is a subset of unsupervised learning, where models are trained on tasks that generate their own supervision signal. In other words, the model learns by solving auxiliary tasks that are designed to help it learn useful representations of the data without using explicit labels. In computer vision, self-supervised learning often involves tasks such as predicting the rotation of an image, colourising greyscale images, solving jigsaw puzzles using shuffled image patches, or predicting hidden pixels. These tasks encourage the model to learn meaningful features and representations, which can later be fine-tuned for specific supervised tasks.

Two main strategies exist for self-supervised pre-training ([Prince, 2023](#)):

Contrastive learning: The goal is to associate similar observations and dissociate dissimilar ones. To create such pairs, data must be either augmented or already come in paired form (e.g. X-rays from different views or a patient or temporally ordered data) ([Krishnan et al., 2022](#)). How best to augment the images is an open question.

Generative learning: The goal is to predict portions of the data that were hidden ([Krishnan et al., 2022](#)). One advantage of this approach is that automating the augmentation is easier and can help to make transformers more data-efficient ([Krishnan et al., 2022](#)). For example, GPT-3 ([Brown et al., 2020](#)) was trained to predict the next word correctly. Similarly for images, a model can learn to predict grayed-out parts ([Atito et al., 2022](#)).

With regard to multimodal approaches, a third strategy can be identified (Khan et al., 2021):

Cross-modal learning: The goal is to leverage information from multiple modalities by learning from the correspondence between, e.g. text and image, audio and video, etc. In healthcare, one example can be to train models for detection tasks from X-rays via X-ray and report pairs (Tiu et al., 2022).

One example for a generative self-supervised learning strategy is Group Masked Model Learning (GMML), which is used to extract contextual information from images (Atito et al., 2022). For that, random groups of connected tokens are manipulated, letting the model recover the hidden information from the visible parts using an autoencoder-based reconstruction loss. The authors built on the Vision Transformer (ViT) (Dosovitskiy et al., 2020) encoder architecture and added a light decoder. The authors discuss different options to replace the selected tokens, concluding that combining masking with noise and replacing with plausible visual concepts (from, for example, another image, which they call an ‘alien concept’) works best. Atito et al. (2022) showed that besides good performance, the model also seems to be less resource intensive, helping to build smaller models trained on smaller datasets. In general, as Bommasani et al. (2021) outlined and, for example, Atito et al. (2022) tackled, there is a strong need to train models more efficiently and more quickly due to increasing data sizes and limited budgets.

Regarding multimodal applications, various self-supervised transformers exist that were built for multimodal image/video-audio-text tasks (Akbari et al., 2021; Alayrac et al., 2020). Here, the models were built and pre-trained on multimodal data and could then be applied to downstream tasks that either contain the same modalities (Alayrac et al., 2020) or are single/mixed-modality downstream tasks (Akbari et al., 2021). These models, such as Radford et al. (2021), usually use image-text pairs with strong correlation assumptions, such as image-caption pairs.

A.1.3 Model backbones – convolutional networks

CNNs are widely used for a couple of years now, particularly in computer vision ([Çalli et al., 2021](#)), owing to their hierarchical feature learning, robustness to translations, transfer learning capabilities, architectural advancements and versatility in application across diverse domains.

At the heart of CNNs are convolutional backbones (which are usually followed by fully-connected layers), composed of convolutional, pooling and activation layers that transform input images into feature maps or representations ([Krizhevsky et al., 2012](#); [Prince, 2023](#)) (cf. [Figure 24](#)). Their main role is to extract hierarchical and meaningful features from the input, suitable for tasks such as image classification and object detection or segmentation. These backbones should identify both local and global patterns in images by using convolutional filters, which move across the image to execute element-wise operations. As the image passes through multiple layers of the network, the backbone learns increasingly abstract and complex features. The main idea here is that a matrix, often called kernel, is moved over the input matrix. Each element of the overlapping matrices is multiplied with each other and then all are summed up. By using different kernels, different shape information can be extracted. Pooling layers, such as max-pooling or average pooling, are used to reduce the dimensions of the feature maps, thereby reducing computational requirements. Activation functions, such as ReLU (Rectified Linear Unit), introduce non-linearity into the network, enabling it to learn complex, non-linear relationships in the data ([Prince, 2023](#)).

There are several groundbreaking convolutional networks in the literature, including:

- LeNet ([LeCun et al., 1998](#)): As one of the earliest CNN architectures, developed by Yann LeCun et al., this architecture was often used for handwritten digit recognition.
- AlexNet ([Krizhevsky et al., 2012](#)): Designed by Alex Krizhevsky et al., this architecture brought CNNs to prominence by winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012.

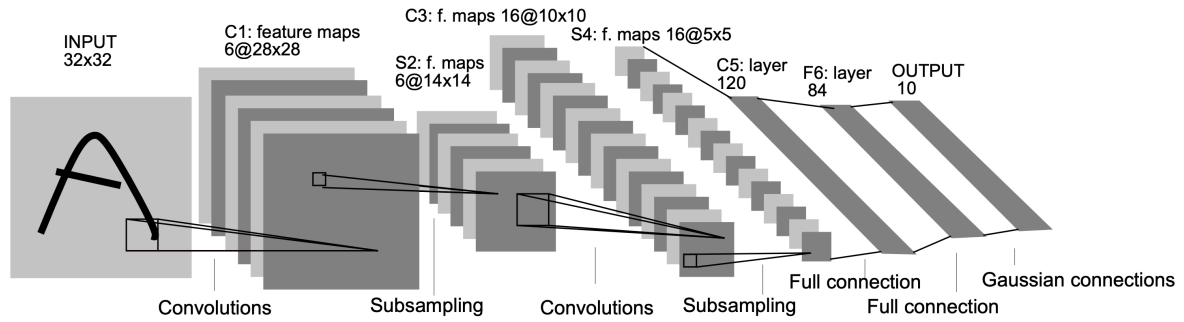


Figure 24: **LeNet-5 CNN**. Architecture by [LeCun et al. \(1998\)](#).

- VGG ([Simonyan and Zisserman, 2014](#)): Developed by the Visual Geometry Group (VGG) at the University of Oxford, this architecture is characterised by its simplicity and the use of small 3x3 convolutional filters.
- ResNet ([He et al., 2016](#)): Introduced by Kaiming He et al., this architecture employs residual connections (skip connections) that help tackle the vanishing gradient problem and allow for deeper networks.
- Inception (GoogLeNet) ([Szegedy et al., 2015](#)): Developed by researchers at Google, this architecture uses a module-based approach with multiple convolutional and pooling layers operating in parallel.
- MobileNet ([Howard et al., 2017](#)): Designed by Andrew G. Howard et al., this architecture focuses on efficient, lightweight networks optimised for devices with fewer computational resources, such as mobile devices.
- DenseNet ([Huang et al., 2017](#)): Developed by Gao Huang et al., this architecture features dense connections between layers, where each layer receives input from all preceding layers, promoting feature reuse and improving gradient flow, which helps in training deeper networks.
- EfficientNet ([Tan and Le, 2019](#)): Introduced by Mingxing Tan and Quoc V. Le, this architecture tries to balance the network width, depth and resolution to be able to uniformly scale up all dimensions. The goal is to improve performance during scale-up, while having fewer parameters and lower computational cost.

A.1.4 Model backbones – transformers

Transformers, built on the self-attention mechanism (Vaswani et al., 2017), do not rely on recurrent or convolutional layers like traditional deep learning models. The self-attention mechanism enables shorter computation paths and allows for the comparison of elements with significant differences across the input data in both directions, as opposed to the unidirectional processing in Recurrent Neural Networks (RNN)s. Unlike the modality-specific nature of convolutional layers or RNNs, transformers offer greater versatility as highlighted in Bommasani et al. (2021). However, their broad applicability requires more data for effective training.

Figure 25 illustrates the original transformer model proposed by Vaswani et al. (2017). The model follows an encoder-decoder architecture, where self-attention modules and fully connected layers are stacked in an N times repeating pattern in both the encoder and decoder. Additionally, Vaswani et al. (2017) incorporated residual connections around both the attention and fully connected layers and introduced normalisation layers in between. The primary distinction between the encoder and decoder is that in the decoder, the attention layer operates over the encoder stack and a masked attention module is added to the output embedding. This design allows transformers to efficiently process and learn from input data, maintaining a higher level of flexibility compared to traditional deep learning models.

At the core of transformers is the self-attention mechanism, which is unique to transformers compared to previous network architectures. It relates different positions of a sequence to compute its representation (Vaswani et al., 2017). It does this by mapping a value vector v , a key vector k and a query vector q to an output vector that represents the weighted sum of the values v , using what Vaswani et al. (2017) call a scaled dot-product attention module¹⁴ (cf. Figure 26). Many interpretations exist for the value, key and query vectors, but essentially, these are just mappings of the raw data using three different weighting matrices. Keys k and queries q are of dimension $d_q = d_k$ and values

¹⁴Instead of performing single attention functions, Vaswani et al. (2017) linearly project everything h times and perform the attention function in parallel.

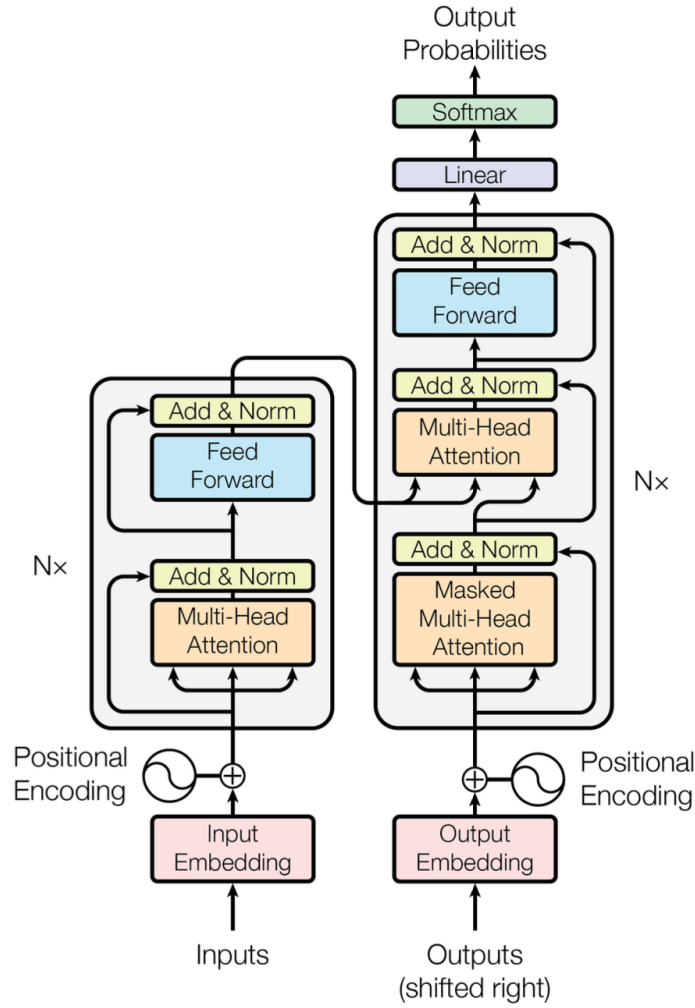


Figure 25: **Transformer architecture.** (Vaswani et al., 2017)

of dimension d_v . The dot product of q and k (scaled by $\frac{1}{\sqrt{d_k}}$) is used as weights on the values v —hence the name ‘attention’, it tells the model what part of the input it should focus on. (For language, this is how relevant the i^{th} word is relative to the other words in the sentence.)

In computer vision, ViT have emerged as a powerful alternative to traditional CNNs for addressing various tasks. The breakthrough was achieved by Dosovitskiy et al. (Dosovitskiy et al., 2020), who first applied the concept of transformers to imaging tasks using supervised learning with minimal modifications to the original structure. In their approach, an input image is divided into non-overlapping fixed-size patches, which are then flattened into 1D vectors and treated as tokens or words, much like in Natural Language Processing (NLP). Each flattened patch is passed through a linear

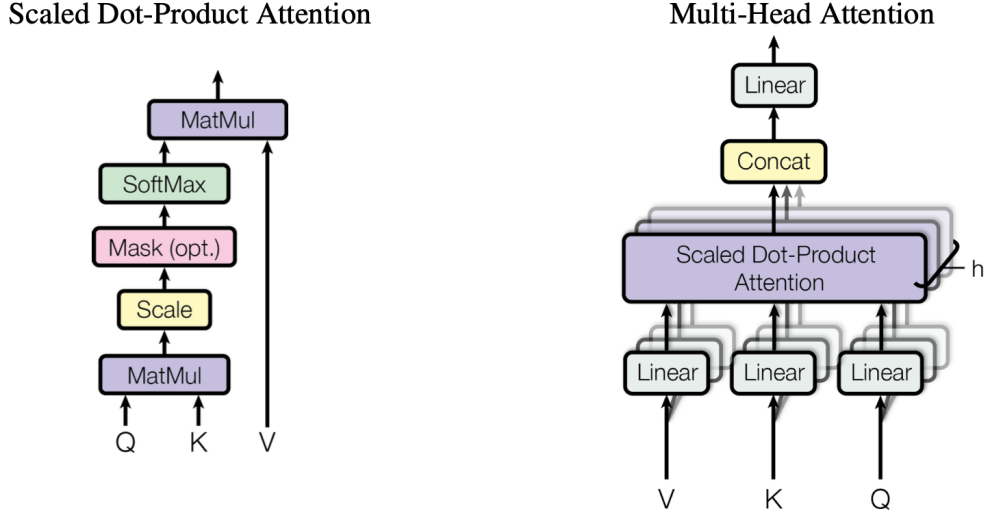


Figure 26: **Attention module.** ‘(left) Scaled dot-product attention. (right) Multi-head attention consists of several attention layers running in parallel.’
(Vaswani et al., 2017)

projection to obtain a fixed-size embedding vector and positional encodings are added to preserve the spatial information lost during patch flattening. The transformer then receives these linear embeddings of patches as input sequence (cf. Figure 27). This architecture allows the model to learn contextual information and global relationships between patches, enabling it to recognize complex patterns and objects in the image. ViTs are typically pre-trained on large datasets and fine-tuned on smaller, task-specific datasets to achieve state-of-the-art performance on various computer vision tasks, such as image classification, object detection and semantic segmentation¹⁵.

One shortcoming of ViT is that due to larger patches, it can struggle with tasks such as object detection or segmentation (Liu et al., 2021). The Swin transformer (Liu et al., 2021) represents a transformer architecture more specifically tailored for imaging tasks. In tasks such as segmentation, the Swin transformer can be beneficial due to its use of smaller patches. A distinctive aspect of Swin transformers is its shifted window design, which bridges connections across windows in consecutive self-attention layers. This allows to use smaller patches and still be computationally efficient. This feature enhances the model’s ability to process visual information. Additionally, its hierarchical

¹⁵Semantic here refers to the idea that the underlying concepts of the images are understood and segmentation does not just happen based on visual or geometric clues.

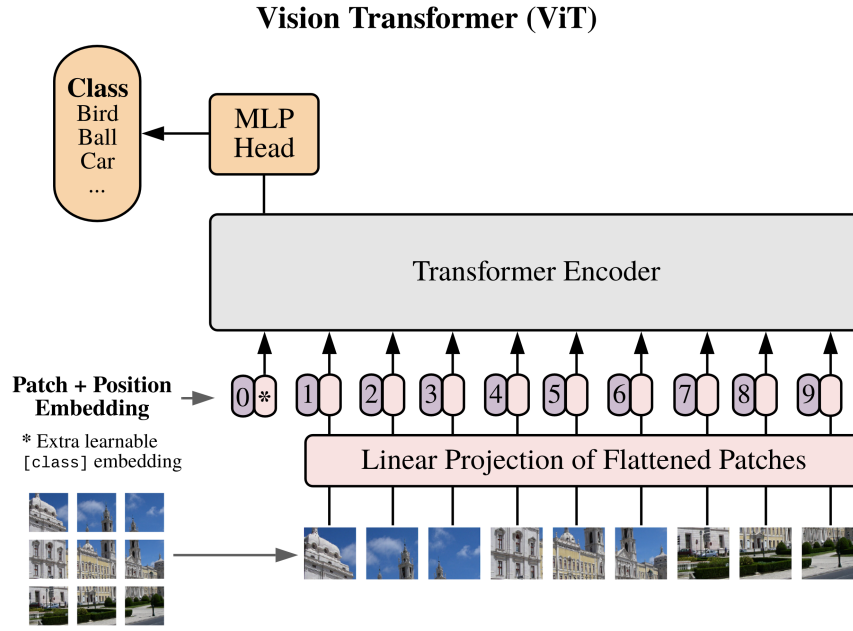


Figure 27: **ViT overview.** ‘An image [is split] into fixed-size patches, linearly embed [...], position embeddings [added] and [...] the resulting sequence of vectors [is fed] to a standard Transformer encoder. [...]’ ([Dosovitskiy et al., 2020](#))

design merges smaller patches into larger ones progressively, enhancing its capability to handle various scales effectively. The architecture is compatible with a broad range of vision tasks and maintains linear computational complexity with respect to image size, making it highly efficient and scalable.

The basis for performance differences between CNNs and ViTs in medical imaging was extensively studied by [Oh et al. \(2024\)](#). The authors showed that while CNNs show better results on clean images and are more invariant towards translation, ViTs show better robustness towards obstructions. On the other side, [Wang et al. \(2022b\)](#) showed that pure CNNs with optimised architectural designs—specifically, patchifying input images, enlarging kernel size and reducing activation and normalisation layers—can achieve robustness on out-of-distribution samples comparable to or exceeding that of Vision Transformers, challenging the assumption that Transformers’ robustness is solely due to self-attention mechanisms.

A.2 AI deployment challenges

This section focuses on current barriers in deploying AI applications in healthcare.

A.2.1 Risks for AI applications in sensitive systems

In sensitive systems, such as healthcare, building machine learning models requires careful attention to certain pitfalls. Such systems can be at the heart of vital decision-making processes that carry substantial implications for individuals, setting distinct challenges for AI integration. [Selbst et al. \(2019\)](#) identified five traps that a technical solution should avoid. The first is *solutionism*, which refers to not recognising that the best solution might not involve technology. The second is the *ripple effect*, where one fails to understand how a new, technical solution changes the behaviour of an existing system. The third trap is *formalism*, meaning not using or understanding the correct definition of fairness in the context. The fourth is *portability*, which involves not correctly understanding the consequences of the re-purposing of an algorithm. Finally, the fifth trap is *framing*, where the entire system is not modelled, meaning the wrong data or labels are chosen.

With respect to the ripple effect, research has shifted towards incorporating humans-in-the-loop for AI applications in sensitive decision-making processes ([De-Arteaga et al., 2020](#)). Understanding how professionals modify their decision-making when using AI systems is essential. Generally, two tendencies are observed, which may vary depending on the context and prior experience with such systems (e.g. past error rates): algorithm aversion and automation bias ([De-Arteaga et al., 2020](#)). Algorithm aversion is the tendency to ignore AI system recommendations, whereas automation bias is the propensity to unquestioningly follow these suggestions. By acknowledging and addressing these tendencies, alongside other challenges, AI systems can be more effectively integrated into sensitive areas, thereby reducing adverse effects.

A.2.2 Deployment challenges in healthcare

Despite substantial progress in clinical AI, many unresolved questions and obstacles remain. To capture this sentiment, Topol (2019) coined the term *AI chasm*, stating: ‘An algorithm with an area under the curve (AUC) score of 0.99 is not worth much if it is not proven to improve clinical outcomes’. Model trust, data constraints and equitable performance across different groups stand out as key challenges in deploying ML in healthcare (Rajpurkar et al., 2022). Model trust is closely associated with model reliability. Model reliability involves aspects like robust generalisation and consistent performance across groups (Glocker et al., 2021; Tran et al., 2022). A reliable model must accurately reflect its own uncertainty, adapt well to new conditions and efficiently process new data (Tran et al., 2022). Despite advancements, generalisation remains an unresolved issue in clinical AI (Azizi et al., 2022). Multi-site evaluation and detecting distribution shifts are essential for generalisation, yet 72% of clinical ML studies neglected multi-site testing (Lu et al., 2021). Lu et al. (2021) show that models often fail to outperform human practitioners in multi-site evaluations and many hence argue for Randomised Control Trial (RCT) as the gold standard in testing applications (Kelly et al., 2019).

Regarding ethical considerations, Morley et al. (2020) outlined six overarching concerns related to AI in healthcare: 1) the inconclusive nature of algorithmic outcomes, 2) the inscrutable nature of algorithmic decisions, 3) the potential for misguided evidence arising from unreliable data, 4) the possibility of unfair outcomes for certain groups, 5) the transformative effects on reality resulting from algorithmic activities and 6) the need for traceability to ensure accountability for any potential harm. Similarly, Ning et al. (2023) summarised eleven ethical principles that need to be considered: accountability, autonomy, beneficence, equity, integrity, morality, non-maleficence, privacy, security, transparency and trust.

Figure 28, by Rajpurkar et al. (2022), provides an overview of the current progress, challenges and opportunities of AI in healthcare.

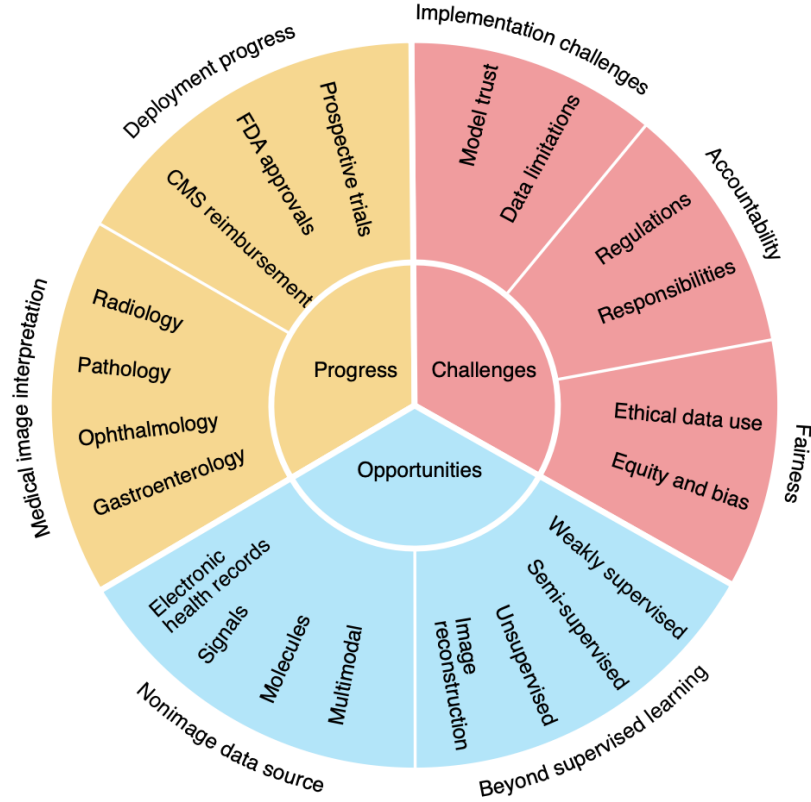


Figure 28: **AI in healthcare.** ‘Overview of the progress, challenges and opportunities for AI in healthcare. CMS, Centers for Medicare & Medicaid Services.’ (Rajpurkar et al., 2022)

A.3 Reliability in medical computer vision

A.3.1 Model reliability

Building AI systems that can operate reliably is the key idea of the push towards more reality-centric AI (Van Der Schaar and Rashbass, 2023); with reliability being defined as ‘the ability for a model to work consistently across real-world settings’ (Tran et al., 2022). With this goal in mind, Tran et al. (2022) derived three main requirements reliable models should fulfil: *Uncertainty*: Models should be able to represent the uncertainty of their predictions. *Generalisability*: Models should show robust performance on new scenarios. *Adaptation*: Models should be able to adapt to new scenarios.

Uncertainty: It is crucial to articulate how a model’s confidence in its predictions correlates with the predictions themselves and how insight into this relationship can

bolster model dependability. Assessing predictive uncertainty allows users to estimate the trustworthiness of the model’s predictions and prompts prudent reliance on its output. Calibration measures, such as expected calibration error, estimate the congruence between a model’s certainty and its precision. Implementing uncertainty quantification aids in informed decision-making; for instance, selective prediction permits a model to defer decisions to human experts when its certainty is low.

Robust generalisation: Understanding model performance on Out-of-Distribution (OOD) data is important for reliable deployment in novel contexts. This facet involves predicting events beyond the model’s training data. While traditional metrics like accuracy and AUC evaluate prediction quality, it is important to also consider robustness to distributional variations, including shifts in covariates or subpopulations.

Adaptation: The model’s ability to adjust to new data swiftly is important. Adaptation refers to the model’s learning efficiency over time and sites. Current benchmarks often assess static datasets with fixed train-test divisions; however, numerous practical applications necessitate models that can rapidly adapt and learn from minimal examples, a trait seen in few-shot learning, active learning and lifelong learning scenarios.

A.3.2 Data drifts

Data drifts present a notable challenge for the application of machine learning models in clinical settings, yet the concept remains ill-defined. [Sahiner et al. \(2023\)](#) characterised data drift in medical imaging as ‘a mismatch between the conditions for model training and clinical use [...] [where] the reference point for the data drift is an ML model that was trained and initially tested with independent data to provide a performance estimate and the mismatch with respect to clinical use may be due to changes of environment and/or changes over time’. [Yoon et al. \(2023\)](#) identified key sources for changes in data distributions. They state variations in imaging equipment, protocols and patient populations, across sites, time and modalities as main drivers (cf. [Figure 29](#)).

Overall, the shifts in data distributions can be categorised into two main concepts

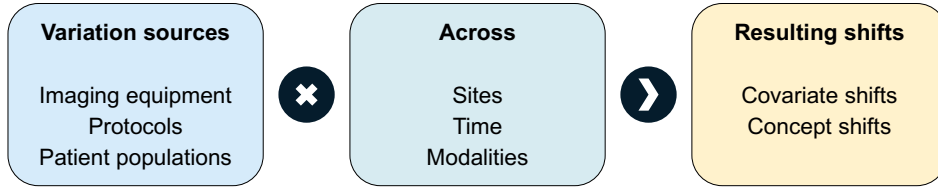


Figure 29: **Data distribution shifts.** Key sources for changes in data distributions in medical imaging, resulting in either covariate shifts or concept shifts.

(Sahiner et al., 2023; Yoon et al., 2023): *Covariate shifts* occur when the distribution of data changes between the source and target domains. Despite this change, the fundamental concept, or the functional connection between inputs (such as images) and outputs (like diagnostic labels), remains unchanged. An example of this could be when a model trained on X-ray images from one hospital encounters different image qualities at another hospital, yet the task of identifying fractures remains the same. *Concept shifts* involve a change in the relationship between the input data and the target outcome. In these shifts, the concept, or the functional link between input and output, undergoes a transformation. For instance, a model trained to identify tumors in brain Magnetic Resonance Imaging (MRI)s might struggle when applied to spinal MRIs, where the tumor characteristics and their implications for diagnosis are different, leading to a change in the underlying relationship between the features and the labels.

A.3.3 Intervention methods

To be able to counteract data shifts, identifying potential data drifts early is important, e.g. by implementing surveillance systems for data distribution shifts (Sahiner et al., 2023). New methodologies are constantly being developed to better ascertain the possible causes of performance degradation (Namkoong et al., 2023).

Solutions to manage data drifts can be roughly categorised into two main categories: *pre-deployment* and *post-deployment* (Sahiner et al., 2023; Yoon et al., 2023).

Pre-deployment solutions operate at three levels. At the *data level*, measures aim to increase diversity and quantity through improved acquisition standards, image manipulations (e.g. normalisation or using surrogates), and image augmentations to generate

new images (e.g. random augmentations, such as flipping and rotation, adversarial-based augmentations, or generative methods). At the *feature level*, measures seek to extract domain-invariant features through improved alignment (e.g. normalisation, such as batch normalisation, or dissimilarity-based alignment) and improved disentanglement. At the *model level*, measures aim to improve model generalisation through improved learning strategies (e.g. ensemble learning, meta-learning, self-supervised learning, or adversarial learning) and improved model structures and optimisations.

Post-deployment solutions focus on improved interpretability and explainability, as well as monitoring and maintenance (e.g. re-training).

A.3.4 Bias and fairness

Bias can arise from imbalanced training data (Mittelstadt et al., 2023; Obermeyer et al., 2019) and manifest in various forms, such as disparate data availability for protected groups or variations in data quality (Mittelstadt et al., 2023). Mehrabi et al. (2021) categorised bias origins into three streams: biases arising from data to algorithms, such as measurement and aggregation biases; from algorithms to users, including algorithmic and presentation biases; and from user to data, which encompasses historical and population biases. The sources of bias are comprehensively summarised at catalogofbias.org.

Discrimination on the other side manifests through decision-making on biased data, taking two primary forms (Wachter et al., 2020): direct discrimination, where treatment varies based on a protected attribute and indirect discrimination, where seemingly neutral policies disproportionately affect a protected group.

Fairness discussions stemming from discussed bias issues often pivot around two principles: disparate treatment and disparate impact. Disparate treatment prohibits decisions based on sensitive attributes, while disparate impact avoids outcomes that disproportionately affect any group (Kleinberg et al., 2016).

When trying to understand subgroup differences better, it is important to clarify how features and tasks can be related. From a machine learning standpoint, there are three

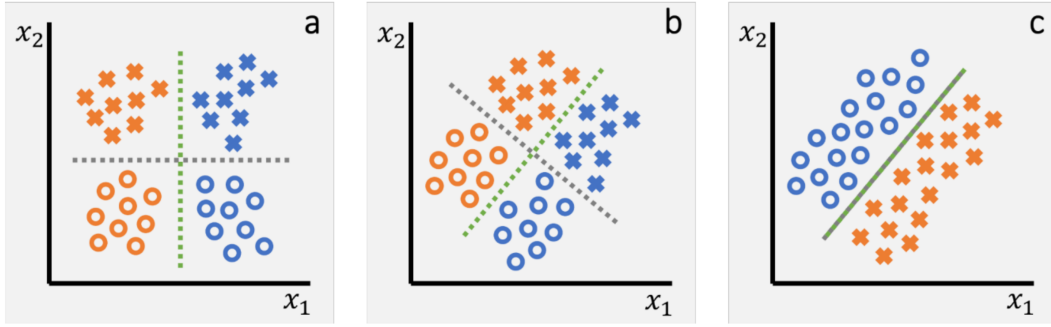


Figure 30: **Bias illustration.** Illustration of different inter-relationships for two classification tasks (colors and shapes) by Glocker et al. (2021): a) Unrelated on feature- and output-level. b) Related on a feature-level but not on outputs. c) Related both on feature- and output-level.

possible scenarios. Using the example from Glocker et al. (2021) (see Figure 30) of separating colors and shapes: In the first scenario, scenario A, the tasks can be solved completely independently, meaning they both use different sets of features. In the second scenario, scenario B, the tasks use to some level the same feature information, but are solved independently of each other, e.g. by applying different weights etc. In scenario C, the tasks themselves are connected, making it seemingly impossible to disentangle them. They use the same features and at least partially the same solving strategies. This means that solving one task will, to some extent, also solve the other task. In practice, various degrees of correlation can exist.

A.3.5 Fairness metrics

Fairness metrics fall into two categories (Wachter et al., 2020): those that preserve bias, ensuring formative equality by replicating historical error rates across groups and those that transform bias, fostering substantive equality by adjusting decision rates to redress historical disparities. An array of fairness metrics is documented, including by Chouldechova (2017), Verma and Rubin (2018) and within the Google fairness glossary¹⁶.

Many metrics are inherently incompatible, each tailored to different facets of fairness. For instance, demographic parity and equal accuracy concern outcomes, whereas others, like predictive equality and equality of opportunity, focus on error rates. The implications

¹⁶<https://developers.google.com/machine-learning/glossary/fairness?hl=en>

of errors, false positives versus false negatives, must be carefully weighed against the context of application.

Two of the most often used metrics are equalised odds and equal opportunity. Equalised odds require the predictor to have equal True Positive Rate (TPR)s for different demographic groups when the outcome is positive and equal False Positive Rate (FPR)s when the outcome is negative¹⁷. Equal opportunity, a relaxed form of equalised odds, only mandates non-discrimination within the advantageous outcome group¹⁸. Both equalised odds and equal opportunity have an advantage over demographic parity, as they do not necessitate independence from the protected attribute. Demographic parity is a fairness metric ensuring equal positive outcome probabilities across different demographic groups, regardless of protected attributes such as race, gender, or age. It aims to prevent discrimination or bias in model predictions.

[Wachter et al. \(2021\)](#) strongly encourage the use of Conditional Demographic Disparity (CDD), ‘as a standard baseline statistical measurement [because it] aligns with [the goals set by] the European Court of Justice’. CDD is a condition test that compares whether for any choice of attributes R and for people with those attributes R , the proportion of protected people in the disadvantaged group is bigger than the proportion of protected people in the advantaged group. That means this condition must hold for all choices of R to classify a system as disparate. For example, let us say we only consider people with the attribute R being the yearly income. The protected attribute is sex. The options are sex A and B , where B is the protected sex. The advantaged group is people who get a loan and the disadvantaged group is people who do not get a loan. Then the ratio of all people who did not get the loan and are from sex B must be bigger for all

¹⁷Example for equalised odds: Suppose a ML model is used to predict whether a person will repay a loan (a binary classification problem where 1 indicates repayment and 0 non-repayment). If we aim to achieve equalised odds, we would want the model to have the same TPR and FPR for all demographic groups. That is, the model should be equally likely to correctly predict repayment for all groups (equal TPR) and also equally likely to incorrectly predict repayment when it is not the case (equal FPR).

¹⁸Example for equal opportunity: Let us take the example of a loan repayment prediction model. Here, to achieve equal opportunity, we would only need to ensure that the model is equally likely to correctly predict loan repayment 1 for all demographic groups. This means the model may have different FPRs for different groups, but it correctly identifies those who will repay the loan at the same rate across all groups. This is particularly important in scenarios where false negatives (failing to identify a positive case) are more harmful than false positives.

groups of yearly incomes than the ratio of all people who got the loan and are from sex B . Because this might be hard to satisfy, I argue that one must also look at the group disparities individually to not miss any injustices.¹⁹

To ensure fair models, particularly in healthcare it is important to examine error rates across subgroups, considering the varying consequences of different errors.

A.3.6 Approaches ensuring fairness

To address the issue of biased outcomes, [Chen et al. \(2021a\)](#) categorised three main approaches for modifying algorithms to ensure fairness: *pre-processing*, *in-processing*, and *post-processing*.

Pre-processing techniques are aimed at mitigating existing biases. These include fairness through unawareness, which avoids the use of sensitive attributes; importance weighting, which adjusts the training distribution to reflect the test distribution; targeted data collection to enhance representation; resampling methods for balancing datasets; data transformation to mitigate bias; and variable blinding to prevent the model from accessing sensitive information.

In-processing techniques are integrated during the learning process. As outlined by Bohdal et al. ([Bohdal et al., 2023](#)), these comprise subgroup balancing by oversampling underrepresented and undersampling overrepresented groups; domain separation, which employs distinct classifiers for different subgroups; adversarial methods designed to produce feature representations that obscure group identities; regularisation and constrained optimisation to discourage fairness metric violations; and disentanglement approaches that separate sensitive attributes from non-sensitive features in data representations.

Post-processing involves calibration techniques to adjust predictive performance parity, ensuring equal ratios of positive predictions to true positives for each subgroup. These

¹⁹Why is it important that the condition holds for each choice of R individually and not only in total? Let us imagine that there is an income group that is almost always rejected and that most people who apply from that group are of sex B . This would mean that, in total, it could look like the ratio of rejected people of sex B is higher than the ratio of accepted people of sex B , but only because the distributions within the different groups of R are not equal.

strategies are designed to integrate fairness into algorithmic development and deployment, addressing biases at various stages of the machine learning pipeline.

A.3.7 Gaps in underdiagnosis research

Failure to diagnose a disease can result in substantial costs for patients, sometimes leading to missed treatment opportunities or even fatality. Therefore, the investigation of underdiagnosis rates has become a focal point in research (Glocker et al., 2021; Seyyed-Kalantari et al., 2021). For example, in their study of the CheXpert dataset (Irvin et al., 2019), Glocker et al. (2021) examined the underdiagnosis rates for various conditions. They adjusted the dataset to balance the representation across subgroups within three categories: ‘No Finding’, ‘Pleural Effusion’ and all ‘Other’ diseases. They reported a 3% increase in the FPR for Black patients under the ‘No Finding’ classification when the test data replicated the training set distribution. This discrepancy diminished after the re-sampling process. However, this approach has considerable limitations:

As shown in Table 19, re-sampling does not guarantee that outcomes will align with ground truth labels (compare experiments 2 and 3). Evaluating the models against the ground truth data²⁰ reveals more significant discrepancies among the subgroups than previously reported. These outcomes are heavily dependent on the data distribution selected for model validation during the training phase (compare experiments 3-7). The ground truth data do not include any Black patients with a positive ‘No Finding’ case (see Table 20); therefore, if models trained and validated on a distribution that mirrors the training data are then tested on the ground truth data, the discrepancy increases, resulting in an elevated FPR for Black patients. Conversely, when a data distribution akin to the ground truth is used for validation, Black patients exhibit a substantially lower FPR for the ‘No Finding’ label.

Moreover, I advocate for an analysis that considers each disease independently. While Glocker et al. (2021) did concentrate on the ‘No Finding’ category, this label is distinct

²⁰The term ‘ground truth’ refers to the 200 manually labelled validation data subsequently provided by the authors of the CheXpert dataset (Irvin et al., 2019).

from others; it cannot overlap with any disease and hence should not be modelled in the same manner as multi-label conditions. In their analysis, [Glocker et al. \(2021\)](#) set a threshold so the FPR for ‘No Finding’ would not exceed 20%. An assessment of their predictive outcomes²¹ indicates numerous instances where the model identified ‘No Finding’ concurrently with another disease.

In conclusion, the analysis showed that performance disparities across subgroups may stem from data distribution variances, not necessarily from inherent model biases. The above shows that discrepancies in underdiagnosis rates are influenced by the composition of test sets. This supports the preferential use of ground truth data over resampling, which is shown to be fraught with limitations. Moreover, I highlighted the importance of a dual-stage approach for the ‘No Finding’ label to ensure interpretability. The findings underscore the risks of indiscriminately applying AI models to new data, stressing the importance of label quality for reliable model performance.

²¹<https://imperialcollegelondon.app.box.com/s/bq87wkuzy14ctsyf8w3hcikwzu8386jj/file/878918991309>

Table 19: **Model performance (AUC) for ‘No Finding’ label, by subgroups.** Trained and evaluated on CheXpert, averaged over five runs with random seeds for training. Based on [Glocker et al. \(2021\)](#)’s code. No patients labelled as ‘Black’ had a positive ‘No Finding’ label in the hand-labelled filtered validation data, hence, there is no AUC. ‘Filtered’ means that the race must contain one of the words ‘Black’, ‘White’ or ‘Asian’ and that ethnicity must be non-Hispanic ([Glocker et al., 2021](#)). Gt = ground truth.

Source	Train/vali set	Test set	Experiment number	White	Asian	Black	Female	Male
AUC (95% CI)								
Glocker et al. (2021)	Glocker et al. (2021)	Glocker et al. (2021)	literature	0.87 (0.86–0.88)	0.88 (0.86–0.89)	0.88 (0.87–0.90)	0.87 (0.86–0.88)	0.87 (0.86–0.88)
Glocker et al. (2021)	Glocker et al. (2021)	Glocker et al. (2021)	literature	0.87 (0.86–0.88)	0.87 (0.87–0.88)	0.89 (0.88–0.89)	0.87 (0.86–0.87)	0.89 (0.88–0.89)
Code run	Glocker et al. (2021)	resampled	1	0.87 (0.86–0.88)	0.87 (0.86–0.88)	0.88 (0.07–0.89)	0.87 (0.87–0.88)	0.88 (0.87–0.88)
Code run	Glocker et al. (2021)	Glocker et al. (2021)	2	0.87 (0.86–0.87)	0.87 (0.87–0.87)	0.89 (0.88–0.89)	0.87 (0.87–0.88)	0.89 (0.88–0.90)
Code run	Glocker et al. (2021)	resampled	3	0.87 (0.87–0.87)	0.92 (0.91–0.92)	na	0.92 (0.92–0.92)	0.84 (0.84–0.84)
Code run	Full train data, train distr. vali data	Gt validation data	4	0.92 (0.91–0.92)	0.92 (0.91–0.93)	0.93 (0.93–0.93)	0.92 (0.92–0.93)	0.92 (0.92–0.92)
Code run	Full train data, gt distr. vali data	Glocker et al. (2021)	5	0.90 (0.90–0.91)	0.91 (0.90–0.92)	0.92 (0.91–0.92)	0.91 (0.90–0.91)	0.91 (0.90–0.91)
Code run	Full train data, gt distr. vali data	Glocker et al. (2021)	6	0.90 (0.90–0.90)	0.90 (0.89–0.90)	0.92 (0.92–0.92)	0.90 (0.90–0.91)	0.91 (0.91–0.91)
Code run	Full train data, gt distr. vali data	resampled	7	0.88 (0.87–0.88)	0.86 (0.86–0.87)	na	0.93 (0.92–0.93)	0.83 (0.83–0.84)
Code run	Full train data, gt distr. vali data	Gt validation data filtered						
TPR (95% CI)								
Glocker et al. (2021)	Glocker et al. (2021)	Glocker et al. (2021)	literature	0.79 (0.77–0.80)	0.80 (0.76–0.83)	0.84 (0.80–0.88)	0.79 (0.77–0.82)	0.79 (0.77–0.81)
Glocker et al. (2021)	Glocker et al. (2021)	Glocker et al. (2021)	literature	0.80 (0.78–0.81)	0.79 (0.78–0.81)	0.81 (0.80–0.82)	0.78 (0.76–0.79)	0.82 (0.81–0.83)
Code run	Glocker et al. (2021)	resampled	1	0.80 (0.79–0.80)	0.81 (0.80–0.81)	0.87 (0.84–0.88)	0.80 (0.79–0.80)	0.81 (0.80–0.81)
Code run	Glocker et al. (2021)	Glocker et al. (2021)	2	0.80 (0.80–0.80)	0.81 (0.80–0.81)	0.84 (0.84–0.85)	0.80 (0.79–0.80)	0.83 (0.82–0.83)
Code run	Glocker et al. (2021)	resampled	3	0.69 (0.69–0.69)	0.75 (0.75–0.76)	na	1.00 (1.00–1.00)	0.58 (0.58–0.59)
Code run	Full train data, train distr. vali data	Gt validation data	4	0.89 (0.89–0.89)	0.91 (0.91–0.91)	0.94 (0.93–0.95)	0.90 (0.90–0.91)	0.89 (0.88–0.90)
Code run	Full train data, gt distr. vali data	Glocker et al. (2021)	5	0.85 (0.84–0.85)	0.85 (0.85–0.86)	0.91 (0.91–0.91)	0.86 (0.85–0.87)	0.86 (0.85–0.87)
Code run	Full train data, gt distr. vali data	Glocker et al. (2021)	6	0.83 (0.82–0.84)	0.85 (0.85–0.86)	0.89 (0.89–0.90)	0.83 (0.83–0.84)	0.87 (0.87–0.87)
Code run	Full train data, gt distr. vali data	resampled	7	0.69 (0.68–0.70)	0.75 (0.75–0.75)	na	1.00 (0.99–1.01)	0.58 (0.58–0.58)
Code run	Full train data, gt distr. vali data	Gt validation data filtered						
FPR (95% CI)								
Glocker et al. (2021)	Glocker et al. (2021)	Glocker et al. (2021)	literature	0.20 (0.20–0.20)	0.20 (0.19–0.21)	0.23 (0.21–0.24)	0.20 (0.20–0.21)	0.20 (0.20–0.20)
Glocker et al. (2021)	Glocker et al. (2021)	Glocker et al. (2021)	literature	0.20 (0.20–0.21)	0.20 (0.20–0.20)	0.20 (0.19–0.20)	0.20 (0.20–0.20)	0.20 (0.20–0.20)
Code run	Glocker et al. (2021)	resampled	1	0.20 (0.20–0.21)	0.19 (0.19–0.20)	0.24 (0.22–0.24)	0.20 (0.20–0.20)	0.20 (0.20–0.20)
Code run	Glocker et al. (2021)	Glocker et al. (2021)	2	0.20 (0.20–0.20)	0.19 (0.19–0.19)	0.21 (0.21–0.21)	0.20 (0.20–0.20)	0.20 (0.20–0.20)
Code run	Glocker et al. (2021)	resampled	3	0.20 (0.20–0.20)	0.22 (0.22–0.22)	0.25 (0.25–0.25)	0.13 (0.13–0.13)	0.26 (0.26–0.26)
Code run	Full train data, train distr. vali data	Gt validation data	4	0.19 (0.19–0.19)	0.20 (0.20–0.20)	0.25 (0.25–0.25)	0.21 (0.21–0.21)	0.19 (0.19–0.19)
Code run	Full train data, gt distr. vali data	Glocker et al. (2021)	5	0.20 (0.20–0.20)	0.19 (0.19–0.19)	0.24 (0.24–0.24)	0.20 (0.20–0.20)	0.20 (0.20–0.20)
Code run	Full train data, gt distr. vali data	Glocker et al. (2021)	6	0.20 (0.20–0.20)	0.19 (0.19–0.19)	0.21 (0.21–0.21)	0.20 (0.20–0.20)	0.20 (0.20–0.20)
Code run	Full train data, gt distr. vali data	resampled	7	0.20 (0.20–0.20)	0.17 (0.17–0.17)	0.12 (0.12–0.12)	0.15 (0.15–0.15)	0.21 (0.21–0.21)
Code run	Full train data, gt distr. vali data	Gt validation data filtered						

Table 20: **Test data distribution of CheXpert**. In *italic*: five labels used in most ML literature. In **bold**: most reported metrics. ‘Filtered’ means that the race must contain one of the words ‘Black’, ‘White’ or ‘Asian’ and that ethnicity must be non-hispanic (Glocker et al., 2021).

	Glocker et al. (2021) split filtered			Glocker et al. (2021) split filtered and resampled				Full split		Full split filtered		
	All	White	Black	Asian	All	White	Black	Asian	All	White	Black	Asian
Test data												
Patients [No.]	12,866	9,956	879	2,031	9,376	6,518	877	1,981	200	112	8	24
Scans [No.]	38,240	29,844	2,746	5,650	114,720	38,605	38,106	38,009	234	133	9	26
Age mean [years]	63	64	57	61	63	63	63	63	61	62	54	62
Age min [years]	18	18	18	18	18	18	18	18	18	18	19	29
Age max [years]	90	90	90	90	90	90	90	90	90	90	87	90
Female [% of patients]	42	41	51	44	46	40	54	44	45	38	56	46
Atelectasis [% of scans]	15	16	13	14	14	15	14	14	34	41	22	35
Cardiomegaly [% of scans]	13	12	21	13	15	12	20	14	29	30	33	35
Consolidation [% of scans]	7	6	6	7	7	7	6	7	14	16	11	15
Edema [% of scans]	26	26	28	23	26	27	29	23	19	19	56	19
Enlarged Cardiome-diast. [% of scans]	5	5	5	5	5	5	5	5	47	47	56	58
Fracture [% of scans]	4	4	2	3	3	4	2	3	0	0	0	0
Lung Lesion [% of scans]	4	4	4	5	4	4	4	5	0	1	0	0
Lung Opacity [% of scans]	51	51	47	52	50	50	49	52	54	59	78	54
Pleural Effusion [% of scans]	41	41	33	44	40	40	41	40	29	29	11	35
Pleural Other [% of scans]	1	1	1	2	1	1	1	1	0	0	0	4
Pneumonia [% of scans]	3	2	3	3	3	2	3	3	3	3	0	4
Pneumothorax [% of scans]	9	9	5	11	8	9	5	11	3	5	11	0
No Finding [% of scans]	9	8	11	9	9	9	9	9	16	17	0	23
Support device [% of scans]	55	56	51	55	54	56	51	55	46	45	44	42
									50	50	50	50

B Lessons learned from multiple years contributing to PhysioNet

Having participated in three consecutive PhysioNet Challenges (2022–2024), I use this Appendix to reflect on the lessons learned from these experiences.

Over recent years, the intersection of machine learning and healthcare has emerged as an active research area, driven by the promise of improving diagnostic accuracy, patient outcomes and healthcare efficiency. Within this field, the annual PhysioNet Challenges have played a significant role, inviting researchers to apply machine learning techniques to pressing healthcare problems. Over the past 25 years, these Challenges have not only showcased the potential of machine learning in healthcare but also highlighted the evolving complexities of clinical data analysis. Recent years reflected challenges across the entire analytics pipeline, from predictive tasks using relatively clean datasets in 2022 (Reyna et al., 2022), to complex scenarios involving long and noisy data in 2023 (Reyna et al., 2023) and ultimately to the 2024 Challenge (Reyna et al., 2024b) that required digitising data from raw inputs. This progression mirrors the broader development of AI in healthcare, moving from idealised laboratory conditions to the complexities of real-world clinical environments.

The transition from lab-based models to practical applications has revealed critical challenges, particularly the gap between model performance on curated datasets and their effectiveness in clinical settings. This discrepancy aligns with calls for more reality-centric and data-centric AI approaches (Van Der Schaar and Rashbass, 2023). As noted by Van Der Schaar and Rashbass (2023), there is a pressing need for methodologies prioritising real-world applicability and resilience to data diversity and quality. The PhysioNet Challenges exemplify this evolving landscape, emphasising the importance of adaptable, robust machine learning models that can handle heterogeneous clinical data.

Contents

B.1	2022: Transfer learning for heart sound classification	194
B.2	2023: Model robustness for ECG classification	194
B.3	2024: Data digitisation for ECG classification	195
B.4	Overarching Challenge lessons	195

B.1 2022: Transfer learning for heart sound classification

The 2022 PhysioNet Challenge (Reyna et al., 2022) focused on developing cost-effective pre-screening tools to detect heart murmurs from heart sound recordings and to determine normality or abnormality of heart sounds. Key lessons included:

Transfer learning as a powerful tool: Converting heart sound recordings into image representations enabled the use of pre-trained image models, improving accuracy and demonstrating adaptability across data representations.

Machine learning’s role in cost-efficient pre-screening: The challenge underscored machine learning’s potential in providing valuable pre-screening solutions, particularly in settings with limited access to expert medical analysis, contributing to global health equity.

Data quality as a persistent hurdle: Comparison of secondary task outcomes with previous Challenges (e.g. Clifford et al. (2016)) and other datasets (e.g. Yaseen et al. (2018)) revealed that cleaner, shorter recordings led to better results, emphasising the need for improving data quality, especially in low-resource settings.

The need for better generalisability: Evaluation of models across datasets (2022, 2016 and Yaseen) revealed significant performance drops when applied outside their original context (Krones and Walker, 2024). This highlights the importance of developing models that are both effective and adaptable across different environments and datasets.

B.2 2023: Model robustness for ECG classification

The 2023 PhysioNet Challenge (Reyna et al., 2023) tasked participants with predicting neurological recovery from coma after cardiac arrest using extended ECG and EEG recordings of varying quality. Key lessons included:

Reevaluating AUC as a standalone metric: While high AUC values were reported, they

did not fully capture practical effectiveness in clinical settings. The challenge scoring highlighted the need for diverse evaluation metrics aligned with real-world healthcare applications.

Leveraging multimodal data: The integration of different data types can improve model performance, but it introduces additional complexity in data management and model design. This emphasises the need for sophisticated integration techniques and high-quality, diverse data sources.

The importance of robust feature engineering: Extracting meaningful features from extensive and noisy ECG data required a deep understanding of both data and clinical phenomena. The success of the winning team ([Zabihi et al., 2023](#)) demonstrated the value of domain knowledge in navigating these complexities.

B.3 2024: Data digitisation for ECG classification

The 2024 PhysioNet Challenge ([Reyna et al., 2024b](#)) focused on digitising ECG recordings from paper and classifying heart diseases, addressing inclusivity and applicability of machine learning in low- and middle-income countries:

The challenge of inclusivity in data sources: Preparing robust data from diverse clinical environments highlighted the complexities of accommodating global variations in data quality and format (notably paper-based ECGs). This underlined the need for comprehensive data preparation strategies to ensure accurate representation and usability for machine learning models.

B.4 Overarching Challenge lessons

The cumulative lessons from these Challenges lead to several key focus areas for future research and development:

Realistic data: Using data that reflect real-world clinical scenarios, including diverse sources and populations, is important. Incorporating OOD data during training and

testing helps to improve model robustness and generalisability (Azizi et al., 2022; Krones and Walker, 2024; Krones et al., 2023; Lu et al., 2021; Reyna et al., 2022, 2023; Van Der Schaar and Rashbass, 2023).

Task-appropriate evaluation metrics: Metrics must reflect practical healthcare applications and consider threshold-dependent error rates, uncertainty quantification, robustness and fairness (Hosseinzadeh Taher et al., 2021; Krones and Walker, 2024; Ma et al., 2022; Pang et al., 2022; Reinke et al., 2024; Tran et al., 2022; Van Der Schaar and Rashbass, 2023).

Robust data preparation and feature engineering: High-quality preprocessing and expert-driven feature extraction are crucial for reliable machine learning models. These steps were highlighted by the 2023 Challenge (Krones et al., 2023; Reyna et al., 2023; Zabihi et al., 2023), where the most successful models incorporated extensive feature engineering, improving accuracy and interpretability.

These insights highlight the ongoing evolution of machine learning in healthcare, from technical advancements to addressing the practical challenges of real-world implementation. Future efforts must prioritise inclusivity, robustness and practical relevance to fully harness machine learning’s potential in healthcare.

C Supplementary material to Chapter 2

This section contains figures and tables with deep-dive information for Chapter 2.

Table 21: **Multimodal healthcare datasets.** We identified more than one dozen publicly available datasets that include multimodal healthcare data.

Dataset name	Use case	Origin	No. observations	Time period	Image data	Text data	Timeseries data	Tabular data	Link and source
MOST	Bones – Osteoarthritis studies	University of Alabama at Birmingham, University of Iowa	> 4,000 patients	2003 – 2010	MRI, X-ray	Reports		Demographics, risk factors, examination results	https://most.ucsf.edu/ (Segal et al., 2013)
OAI	Bones – Osteoarthritis studies	Various sites across North America	> 26M images from 4,796 patients	2004 – 2006	MRI, X-ray			Demographics, clinical data, examination results	https://nda.nih.gov/oai/ (Nevitt et al., 2006)
ABIDE I + II	Brain – Autism studies	Various sites across North America	> 2,000 patients	na – 2017	MRI			Demographics, phenotypic info, intelligence scores, examination scores	http://fcon.1000.projects.nitrc.org/indi/abide/ (Di Martino et al., 2014)
ADNI I - III	Brain – Dementia detection	Various sites across North America	4K MRI, 2K fMRI, 1.5K DTI, 2.5K PET images from 5K patients	2004 – 2022	MRI, PET	Reports	Vital signs	Demographics, statistics from images, status info, longitudinal outcomes	https://adni.loni.usc.edu/ (Weiner et al., 2017)
OASIS-3	Brain – Dementia detection	Washington University Knight Alzheimer Disease Research Center	2,842 MIR, 1,472 CT, 2,157 PET images from 1,379 patients	> 30 years before 2019	MRI, CT, PET			Demographics, statistics from images, status info, longitudinal outcomes	https://www.oasis-brains.org/ (LaMontagne et al., 2019)
IDA (archive)	Brain – Neuroscience disease prediction	Various studies	143 85,007 patients in total	Various	MRI, CT, PET, EEG, SPECT			Demographics, genetics, exam scores	https://ida.loni.usc.edu/login.jsp
Ischemic Stroke	Brain – Stroke treatment outcome	University of Heidelberg	246 patients	2014 – 2018	CT			Demographics, clinical scores	(Brugnara et al., 2020)
BCDR	Breast – Cancer analysis	Centro Hospitalar São João, at University of Porto	1,734 patients	1994 – 2009	Mammograms, Ultrasound images	Reports		Demographics, clinical history, segmentations	https://bcdre.eu/
DDSM (archive)	Breast – Cancer analysis	Various sites across North America	2,620 studies	1988 – 1999	Mammograms			Demographics, breast density rating, abnormal ratings	http://www.eng.usf.edu/cvprg/Mammography/Database.html
TCIA (archive)	Cancer – Archive of datasets	Various studies	180 1 - 26,254	Various	MRI, CT, PET, Pathology			Demographics, outcomes, genetics, treatment details	https://www.cancerimagingarchive.net/
TCGA (archive)	Cancer – Cancer analysis	Various sites across North America	11,007 images from 86,513 patients	2006 – today	Histopathology	Reports		Demographics, omics, clinical data, examination results	https://portal.gdc.cancer.gov/
CheXpert	Chest – Image exploration	Stanford Hospital	224,316 images from 65,240 patients	2002 – 2017	X-ray			Demographics	https://stanfordmlgroup.github.io/competitions/chexpert/ (Irvin et al., 2019)
MIMIC-IV	Chest – Critical care data	Beth Israel Deaconess Medical Center emergency department	69,619 unique ICU stays for 50,048 patients, 361,363 images from 64,586 patients	2008 – 2019, 2011 – 2017	X-ray	Reports	Vital signs, fluid inputs, etc.	Demographics, tracking data, EHR	https://physionet.org/content/mimiciv/2.2/ (Johnson et al., 2019b, 2020)
NIH Chest X-ray	Chest – Image exploration	NIH Clinical Center Bethesda, MD, USA	108,948 images from 32,717 patients	1992 – 2015	X-ray			Demographics	https://www.kaggle.com/datasets/nih-chest-xrays/data (Wang et al., 2017)
PADCHEST	Chest – Image exploration	Hospital San Juan (Spain)	160,868 images from 69,882 patients	2009 – 2017	X-ray	Reports		Demographics	https://bimcv.cipf.es/bimcv-projects/padchest/ (Bustos et al., 2020)
HAM10000	Skin – Lesions detection	Medical University of Vienna, Cliff Rosendahl in Queensland	10,015 images	1998 – 2018	Dermatoscopic images			Demographics	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T (Tschandl et al., 2018)
Biobank (archive)	Various – Disease prediction	United Kingdom	500K patients	2006 – present	MRI, Ultrasound, X-ray			Demographics, genetics, exam scores, biomarkers, survey data, outcome data	https://www.ukbiobank.ac.uk/

Table 22: **Studies using multimodality approaches in brain disorder prediction.** We describe the fusion approach during fine-tuning, not pre-training; Early fusion can also involve heavy feature extraction using ML methods, the difference to intermediate fusion is that the loss is not backpropagated; While we only listed MRI as one modality, MRI can itself stand for various MRI modalities; ‘NN’ as ML methods stands for various neural network architectures outside of the classical RNN, AE or CNN architectures; ‘Feature extraction’ stands for various methods of feature extraction outside the mentioned architectures; GCN = Graph Convolutional Network; AE = Autoencoder; RF = Random Forrest; GB = Gradient Boosting.

Topic	Study, Year	Training strategy	Fusion technique	Image modalities	Timeseries modalities	Text modalities	Tabular modalities	ML methods before fusion	ML methods after/for fusion	No. observations
Alzheimer	Parisot et al. (2018), 2018	Semi-supervised	Early	MRI			Demographics	Feature extraction	GCN	> 2,000
Alzheimer	Huang and Chung (2020), 2020	Semi-supervised	Intermediate	MRI			Demographics, genetic data	GCN	GCN	1,675
Alzheimer	Li and Fan (2019), 2019	Supervised	Early	MRI	Cognitive measurements		Demographics, assessment scores	RNN, feature extraction	Cox regression	822
Alzheimer	Dwivedi et al. (2022), 2022	Supervised	Early	MRI, PET				Feature extraction, CNN	SVM	200
Alzheimer	Zhou et al. (2019), 2019	Supervised	Intermediate	MRI, PET			Risk factors from genetic data	Feature traction	NN	805
Alzheimer	Thung et al. (2017), 2017	Supervised	Intermediate	MRI, PET			Demographics, genetic data	Feature traction	NN	805
Alzheimer	El-Sappagh et al. (2020), 2020	Supervised	Intermediate	MRI, PET	Vital signs		Demographics, assessment scores	Feature traction, CNN, RNN, NN	NN	1,536
Alzheimer	Suk et al. (2014), 2014	Supervised	Intermediate	MRI, PET				NN DBM	SVM	398
Alzheimer	Spasov et al. (2018), 2018	Supervised	Intermediate	MRI			Demographics, genetic data, assessment scores	CNN, NN	NN	376
Alzheimer	Pölsterl et al. (2021), 2021	Supervised	Intermediate	MRI			Demographics, clinical measurements	CNN	NN	1,341 + 755
Alzheimer	Venugopalan et al. (2021), 2021	Unsupervised Supervised	Various	MRI			Demographics, clinical measurements, genetics	CNN, AE	NN	2,004
Bipolar disorder	Achalia et al. (2020), 2020	Supervised	Early	MRI			Demographics, neurocognitive measures	Feature traction	SVM	60
Bipolar disorder	Ceccarelli and Mahmoud (2022), 2022	Supervised	Late	Video	Audio recording	Transcripts		RNN	RF	> 150
Mild cognitive impairment	Qiu et al. (2018), 2018	Supervised	Late	MRI			Demographics, assessment scores	CNN, NN	Majority voting	386
Multiple sclerosis conversion	Yoo et al. (2019), 2019	Unsupervised	Intermediate	MRI			Demographics, clinical measurements	Feature traction, DBN, CNN	NN	140
Schizophrenia	Ghosal et al. (2021), 2021	Unsupervised	Intermediate	MRI			Genetic data	AE	NN	> 100
Coma outcome prediction	Zheng et al. (2021), 2021	Supervised	Late		EEG		Demographics	CNNs, RNNs	Majority voting	1,038

Table 23: **Studies using multimodality approaches in cancer prediction.** We describe the fusion approach during fine-tuning, not pre-training; Early fusion can also involve heavy feature extraction using ML methods, the difference to intermediate fusion is that the loss is not backpropagated; While we only listed MRI as one modality, MRI can itself stand for various MRI modalities; ‘NN’ as ML methods stands for various neural network architectures outside of the classical RNN, AE or CNN architectures; ‘Feature extraction’ stands for various methods of feature extraction outside the mentioned architectures; GCN = Graph Convolutional Network; AE = Autoencoder; RF = Random Forrest; GB = Gradient Boosting.

Topic	Study, Year	Training strategy	Fusion technique	Image modalities	Timeseries modalities	Text modalities	Tabular modalities	ML methods before fusion	ML methods after/for fusion	No. observations
Brain tumor survival prediction	Nie et al. (2019), 2019	Supervised	Early	MRI			Demographics	CNN	SVM	68
Brain tumor survival prediction	Braman et al. (2021), 2021	Supervised	Intermediate	MRI, Pathological images			Demographics, clinical measurements, DNA results	CNN, NN	NN	176
Breast cancer chemotherapy response prediction	Duanmu et al. (2020), 2020	Supervised	Intermediate	MRI			Demographics, genomic data	CNN, NN	NN	112
Breast cancer	Yala et al. (2019), 2019	Supervised	Intermediate	X-ray			Demographics, risk factors	CNN	NN	71M examinations of 89K patients
Breast cancer	Yan et al. (2021), 2021	Unsupervised	Intermediate	Pathological images			Structured data from EMR	AE, CNN	NN	185
Breast cancer	Liu and Hu (2019), 2019	Unsupervised	Intermediate		Genes		Demographics, genomic data	AE	AE	24K genes for 1K patients
Breast cancer survival prediction	Li et al. (2020a), 2020	Unsupervised	Intermediate	Pathological images			Genomic data	CNN, AE	AE	826
Breast cancer	Holste et al. (2021), 2021	Supervised	Various	MRI			Demographics, clinical indications	CNN, NN	NN	5,248
Carcinoma	Kharazmi et al. (2018), 2018	Unsupervised	Early	Dermatological images			Demographics	AE	Softmax classifier	1,191
Lung cancer diagnosis	Hyun et al. (2019), 2019	Supervised	Early	PET, CT			Demographics	Feature extraction	Various	396
Lung cancer response prediction to immunotherapy	Vanguri et al. (2022), 2022	Supervised	Early	CT, Pathological images			Demographics, genomic data	Feature extraction	NN	247
Pan-cancer prognosis	Silva and Rohr (2020), 2020	Supervised	Early	Pathological images			Demographics, clinical measurements, DNA results	Feature extraction, CNN, NN	NN	11,081
Pan-cancer prognosis	Cheerla and Gevaert (2019), 2019	Unsupervised	Intermediate	Pathological images			Demographics, genetics	NN, CNN	Siamese network	11,160
Prostate cancer detection	Rubinstein et al. (2019), 2019	Unsupervised	Intermediate	PET, CT				AE	AE	
Prostate cancer detection	Reda et al. (2018), 2018	Unsupervised	Late	MRI			Biomarkers	Feature extraction, AE	AE	18
Renal cancer prognosis prediction	Schulz et al. (2021), 2021	Supervised	Intermediate	MRI, CT, Pathological images			Genomic data	CNN, NN	NN	230
Tumor soft-tissue sarcoma segmentation	Guo et al. (2019), 2019	Supervised	Various	MRI, PET, CT				CNN	NN or RF	50
Various cancer	Chen et al. (2020), 2020	Unsupervised Supervised	Early	Pathological images			DNA results	GCN, CNN, NN	Various	> 2,000 samples

Table 24: **Studies using multimodality approaches in predicting chest related conditions.** We describe the fusion approach during fine-tuning, not pre-training; Early fusion can also involve heavy feature extraction using ML methods, the difference to intermediate fusion is that the loss is not backpropagated; While we only listed MRI as one modality, MRI can itself stand for various MRI modalities; NN as ML methods stands for various neural network architectures outside of the classical RNN, AE or CNN architectures; ‘Feature extraction’ stands for various methods of feature extraction outside the mentioned architectures; GCN = Graph Convolutional Network; AE = Autoencoder; RF = Random Forrest; GB = Gradient Boosting.

Topic	Study, Year	Training strategy	Fusion technique	Image modalities	Timeseries modalities	Text modalities	Tabular modalities	ML methods before fusion	ML methods after/for fusion	No. observations
Cardiomegaly	Palepu and Beam (2022) , 2022	Self-supervised	Intermediate	X-ray		Reports		Transformer	Transformer	> 200,000
Cardiomegaly	Duvieusart et al. (2022) , 2022	Supervised	Intermediate and Early	X-ray	Vital signs	Reports	Demographics, clinical measurements	Feature traction, CNN	NN	2,774
Cardiovascular risk prediction	Bagheri et al. (2020) , 2020	Supervised	Early			Reports	Demographics, clinical measurements	RNN, FNN	NN	5,603
Cardiomegaly	Grant et al. (2021) , 2021	Supervised	Intermediate	X-ray	Vital signs		Demographics	Feature traction, CNN	NN	2,571
Cardiomegaly	Baltruschat et al. (2019) , 2019	Supervised	Intermediate	X-ray			Demographics	Feature traction, CNN	NN	6,245
Endovascular treatment outcome	Brugnara et al. (2020) , 2020	Supervised	Early	CT			Assessment scores	Feature traction	GB	246
Endovascular treatment outcome	Samak et al. (2020) , 2020	Supervised	Intermediate	CT			Demographics, biomarkers	Feature traction, CNN	NN	500
Heart murmur detection	Walker et al. (2022) , 2022	Supervised	Early		Audio recording		Demographics	Bayesian ResNet	Bayesian ResNet, GB	942
Pathway analysis	Nishimori et al. (2021) , 2021	Supervised	Early	X-ray	ECG			CNN	CNN	294
Pulmonary oedema assessment	Chauhan et al. (2020) , 2020	Semi-supervised	Intermediate	X-ray		Reports		CNN, Transformer	NN	247,425
Covid detection	Xu et al. (2020) , 2020	Supervised	Early	CT			Demographics, clinical measurements, lab results	CNN	Various	689
Covid progression prediction	Fang et al. (2021) , 2021	Supervised	Early	CT			Demographics, clinical measurements	CNN, NN	RNN, NN	1,040
Covid severity prediction	Zhou et al. (2021a) , 2021	Supervised	Late	CT			Demographics, clinical measurements	CNN, NN	NN	733

Table 25: **Studies using multimodality approaches in predicting skin related conditions and other diseases.** We describe the fusion approach during fine-tuning, not pre-training; Early fusion can also involve heavy feature extraction using ML methods, the difference to intermediate fusion is that the loss is not backpropagated; While we only listed MRI as one modality, MRI can itself stand for various MRI modalities; NN as ML methods stands for various neural network architectures outside of the classical RNN, AE or CNN architectures; ‘Feature extraction’ stands for various methods of feature extraction outside the mentioned architectures; GCN = Graph Convolutional Network; AE = Autoencoder; RF = Random Forrest; GB = Gradient Boosting.

Topic	Study, Year	Training strategy	Fusion technique	Image modalities	Timeseries modalities	Text modalities	Tabular modalities	ML methods before fusion	ML methods after/for fusion	No. observations
Image modality combination	Taleb et al. (2021) , 2021	Self-supervised (via puzzle-solving)	Early	MRI, CT				CNN	CNN	> 350
Image recognition	Huang et al. (2021) , 2021	Supervised	N/A	X-ray		Reports		CNN	Transformer	> 200,000
Microcytic hypochromic	Purwar et al. (2020) , 2020	Supervised	Early	Pathological images			Blood test features	CNN	NN	20
Mortality risk for ICU patients	Jin et al. (2018) , 2018	Supervised	Intermediate		Vital signs	Reports		RNN, NN	NN	42,276
Neonatal post-operative pain assessment	Salekin et al. (2021) , 2021	Supervised	Late	Video	Audio recording			CNN, RNN	Majority voting	> 200
Osteoarthritis	Tiulpin et al. (2019) , 2019	Supervised	Early	X-ray		Reports	Demographics, assessment scores	Feature traction, CNN	ex-GB	> 5,000
Various diseases	Rodin et al. (2019) , 2019	Supervised	Intermediate	X-ray		Reports	Demographics	Feature traction, CNN	ex-RNN	8,530
Skin lesions classification	Yap et al. (2018) , 2018	Supervised	Early	Dermatological images			Demographics	Feature traction, CNN	ex-NN	2,917
Skin lesions classification	Gessert et al. (2020) , 2020	Supervised	Intermediate	Dermatological images			Demographics	CNN, NNs	CNN ensemble	25,331
Skin lesions classification	Kawahara et al. (2018) , 2019	Supervised	Intermediate	Dermatological images			Demographics	CNN	NN	1,011

D Supplementary material to Chapter 5

This section includes supporting material such as detailed derivations, extended tables, or additional figures that are directly related to the main text but would interrupt its flow if included.

D.1 Data selection

Table 26: **Label selection for circulatory system disorders.** Refinement of WHO ICD-10 codes in order to derive prediction labels. All prediction groups are modelled as binary prediction problems.

Label group name	ICD-10 blocks	Motive	Description
(1) Hypertensive diseases	I10 - Essential (primary) hypertension I11 - Hypertensive heart disease I12 - Hypertensive renal disease I13 - Hypertensive heart and renal disease	All same underlying disease processes	Long-term, persistent high blood pressure (hypertension) that causes damage to the heart, blood vessels and/or kidneys.
(2) Ischaemic heart diseases	I20 - Angina pectoris I21 - Acute myocardial infarction I22 - Subsequent myocardial infarction I24 - Other acute ischaemic heart diseases	Acute, same underlying processes	Conditions where the heart muscle does not receive enough blood because the coronary arteries are narrowed or blocked, typically leading to chest pain or acute events such as a heart attack.
(3) Chronic ischaemic heart disease	I25 - Chronic ischaemic heart disease	Chronic, same underlying processes	Persistent or long-standing reduction of blood and oxygen flow to the heart muscle, often resulting in lasting damage, scar tissue, or chronic stable angina.
(4) Cardiomyopathies	I42 - Cardiomyopathy I43 - Cardiomyopathy in diseases classified elsewhere	Grouping of all cardiomyopathies	Primary disorders of the heart muscle (myocardium) that weaken its structure and ability to pump, excluding causes like coronary artery disease or hypertension.
(5) Dysrhythmias	I44 - Atrioventricular and left bundle-branch block I45 - Other conduction disorders I47 - Paroxysmal tachycardia I48 - Atrial fibrillation and flutter I49 - Other cardiac arrhythmias	Grouping of all dysrhythmias	Disorders involving abnormalities of cardiac rhythm or impulse conduction.
(6) Heart failure	I50 - Heart failure	Heart failure	The inability of the heart to effectively pump or fill with enough blood to meet the body's metabolic demands.

Table 27: **Cohort filters.** Filters applied to MIMIC-IV to derive final cohort size of 28,262 patients.

Criteria	Count
All hospitalised patients	223,452
Ambiguous ED admissions	-916
No ICU stay	-157,532
No ICU discharge	-11
Died during first ICU stay	-4,716
First ICU stay < 24 h	-12,291
No ED admission	-19,724
Age < 18	-0
Cohort size	28,262

Table 28: **Demographics.** Statistical summary of demographic characteristics in the cohort.

Feature	Value
Total n	28,262 patients
Race	
Asian	3%
Black	13%
Hispanic	4%
Other	17%
White	63%
Gender	
Female	45%
Male	55%
Age mean (SD)	65.4 (18.6) years
Length of stay mean (SD)	272.8 (284.4) hours

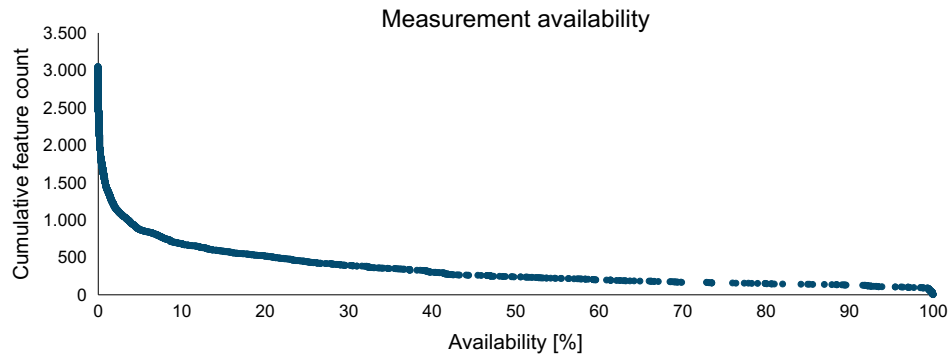


Figure 31: **Measurement availability.** Availability of measures in the cohort is shown, with the y-axis representing the cumulative count of measures based on their availability per patient. For example, 228 measures are available for more than 60% of the patients.

Table 29: **Data selection for MIMIC-IV cohort.** Available features used in modelling. Aggregation refers to the method used for aggregating low-frequency features, where all aggregations were provided to the model as separate features with an option to remove highly correlated features. In bold: Modality type – Sub-type – Category – Location (data source). Availability refers to the availability in the 28k patient cohort. 1/2

Description Feature	Item ID	Unit	Usage information Availability [%]	Aggregation
Tabular - Static patient information (admissions, icustays, patients, medrecon)				
Age	N/A	years	100	at admission
Gender	N/A	categorical	100	N/A
Race	N/A	categorical	80.1	N/A
Prev. medication	N/A	categorical	100	N/A
Tabular - Diagnoses (diagnosis)				
ED ICD codes	N/A	categorical	100	last
Tabular - Low-frequency - Vital signs - ED (vitalsign, triage)				
Diastolic blood pressure	N/A	mmHg	37.3	first, last, mean
Heart rate	N/A	beats per min	37.3	first, last, mean
Order of priority	N/A	score 1-5	36	first, last, mean
Oxygen saturation	N/A	percentage	37.3	first, last, mean
Respiratory rate	N/A	breaths per min	37.3	first, last, mean
Self-reported pain	N/A	score 0-10	37.3	first, last, mean
Systolic blood pressure	N/A	mmHg	37.3	first, last, mean
Temperature	N/A	degrees F	37.3	first, last, mean
Tabular - Low-frequency - Vital signs - ICU (chartevents, oasis, sofa)				
Albumin	227456	g/dl	42.5	mean
Arterial CO2 pressure	220235	mmHg	42.3	mean, max, min
Arterial O2 pressure	220224	mmHg	42.3	mean, max, min
Diastolic blood pressure	220180	mmHg	99.7	mean, max, min
GCS - eye opening	220739	score 1-4	100	last
GCS - motor response	223901	score 1-6	100	last
GCS - verbal response	223900	score 1-5	100	last
Heart rate	220045	beats per min	100	mean, max, min
Inspired O2 fraction	223835	percentage	48.1	mean, min
Mean blood pres.	220181	mmHg	99.7	mean, max, min
Oxygen saturation	220277	percentage	100	mean, min
Platelet count	227457	k/ul	99.4	mean
Respiratory rate	220210	insp/min	99.9	mean, max, min
Sodium (serum)	220645	meq/l	99.5	mean, max, min
Systolic blood pressure	220179	mmHg	99.7	mean, max, min
Temperature	223761	degrees F	99.8	mean, max, min
Weight	226512	kg	100	last
OASIS score	N/A	score	100	first, last, max
SOFA score	N/A	score	100	first, last, max
Tabular - Low-frequency - Treatments - ICU (inputevents)				
Dextrose 5%	220949	ml	64.8	sum, mean
Magnesium sulfate	222011	grams	45.5	sum, mean
NaCl 0.9%	225158	ml	79.4	sum, mean
Po intake	226452	ml	71.8	sum, mean
Solution	225943	ml	48.8	sum, mean
Tabular - Low-frequency - Treatments - Hosp (procedures_icd)				
Procedure codes	N/A	categorical	100	last
Images (metadata)				
X-rays	N/A	N/A	24.4	Last
High-frequency measures (record_list)				
ECGs	N/A	N/A	48.9	Last
Text data (triage)				
ED complaints	N/A	N/A	55.7	Last
Radiology reports	N/A	N/A	75.4	Last

Table 30: **Data selection for MIMIC-IV cohort.** Available features used in modelling. Aggregation refers to the method used for aggregating low-frequency features, where all aggregations were provided to the model as separate features with an option to remove highly correlated features. In bold: Modality type – Sub-type – Category – Location (data source). Availability refers to the availability in the 28k patient cohort. 2/2

Description Feature	Item ID	Unit	Usage information	
			Availability [%]	Aggregation
Tabular - Low-frequency - Lab measures - Hosp (labevents)				
Alanine aminotransferase	50861	iu/l	69.1	mean
Alkaline phosphatase	50863	iu/l	68.5	mean
Anion gap	50868	meq/l	99.7	mean
Bicarbonate	50882	meq/l	99.7	mean
Bilirubin, total	50885	mg/dl	68.6	mean
Calcium, total	50893	mg/dl	99.4	mean
Chloride	50902	meq/l	99.7	mean
Creatinine	50912	mg/dl	99.7	mean
Glucose	50931	mg/dl	99.7	mean
Hematocrit	51221	percentage	99.7	mean
Hemoglobin	51222	g/dl	99.6	mean
INR (PT)	51237	N/A	92.6	mean
Lactate	50813	mmol/l	64.2	mean
Magnesium	50960	mg/dl	99.5	mean
MCH	51248	pg	99.6	mean
MCHC	51249	percentage	99.6	mean
MCV	51250	fl	99.7	mean
Phosphate	50970	mg/dl	99.4	mean
Platelet count	51265	k/ul	99.7	mean
Potassium	50971	meq/l	99.7	mean, max, min
PT	51274	sec	92.6	mean
PTT	51275	sec	92.2	mean
RDW	51277	percentage	99.6	mean
Red blood cells	51279	m/ul	99.7	mean
Sodium	50983	meq/l	99.7	mean
Urea nitrogen	51006	mg/dl	99.7	mean
WBC	51516	num/hpf	47.3	mean, max, min
White blood cells	51301	k/ul	99.6	mean

D.2 Results

Table 31: **Comparison of proposed fusion pipeline to state-of-the-art models.** State-of-the-art models from the literature where trained and evaluated on the proposed classification task the 6 different ICD-10 code groups for the circulatory system (Table 26). Reported is the average (SD) AUC over 5-fold CV. If multiple fusion strategies were proposed, the best reported one was used. For a fair comparison, the fusion dataset of size 5,280 patients was used for all models (Table 11). In bold the best performing one.

Paper/Model	(1)	(2)	(3)	(4)	(5)	(6)
Baseline (Late, fusion, w/o notes)	.757 (.012)	.767 (.016)	.784 (.012)	.732 (.033)	.792 (.009)	.828 (.007)
HAIM (Soenksen et al., 2022)	.754 (.034)	.700 (.039)	.762 (.027)	.676 (.093)	.808 (.006)	.796 (.012)
MedFuse (Hayat et al., 2022)	.767 (.015)	.760 (.015)	.778 (.023)	.775 (.051)	.757 (.021)	.803 (.021)
AutoPrognosis-M (Imrie et al., 2025)	.755 (.012)	.729 (.041)	.783 (.034)	.719 (.030)	.785 (.027)	.808 (.016)

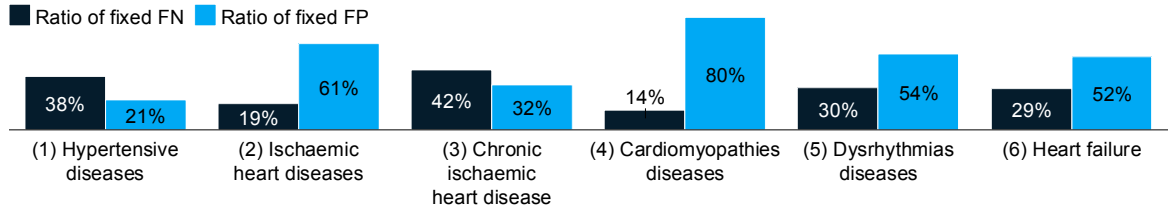


Figure 32: **Fusion effect on incorrect single modality predictions.** The figure shows the effect of the fusion model on incorrectly predicted cases of the main single modality model, considering X-rays, ECGs and LVD data. The figure shows how many false positives and negatives the fusion model fixed.

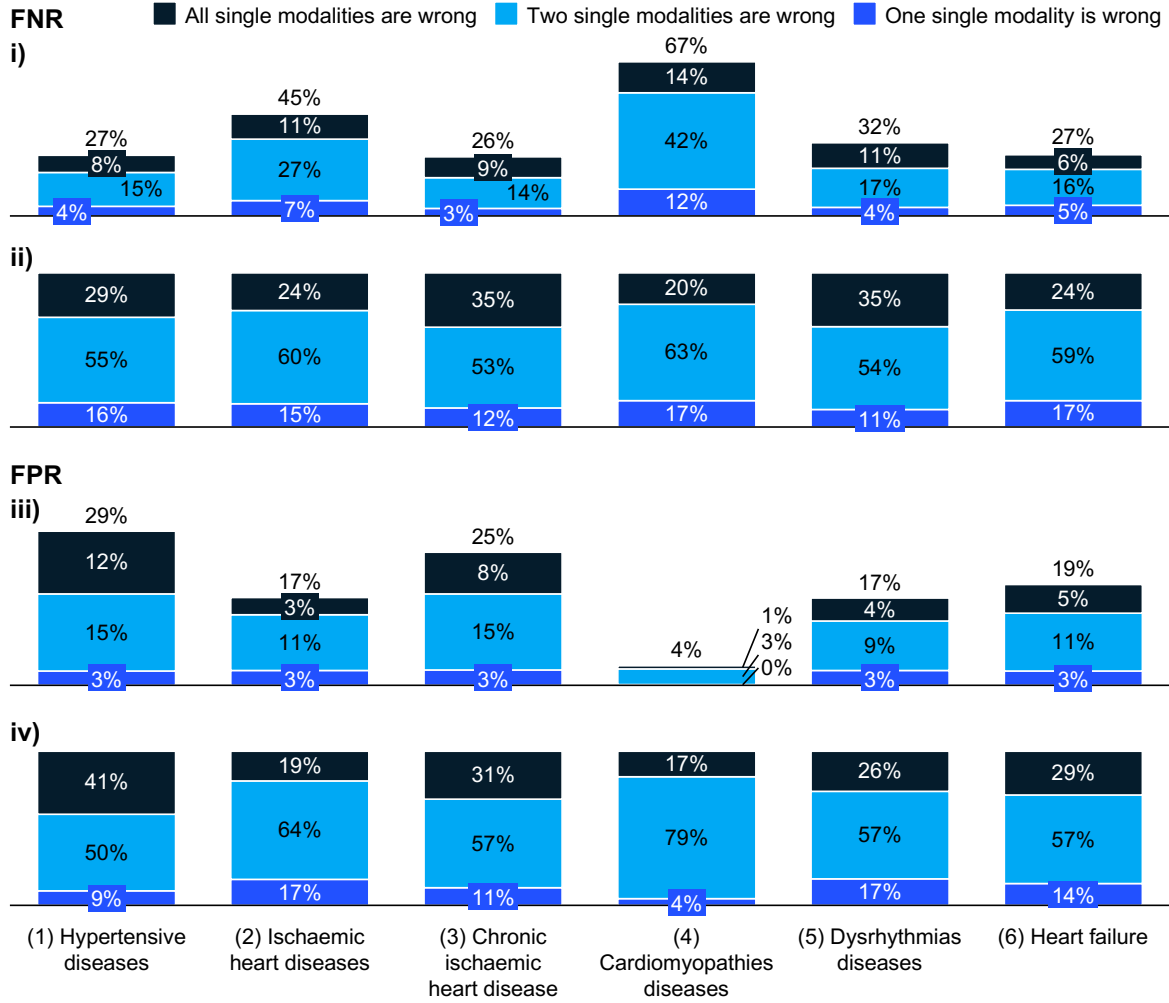


Figure 33: **Modality effect on false fusion predictions.** The figure shows the influence of the single modalities on false predictions of the late fusion model, considering X-rays, ECGs and LVD data. i) and ii) show the split of the FNR into whether all single modalities were wrong, only two or only one, in i) scaled to all predictions and in ii) scaled to the false negative ones. iii) and iv) do the same for the FPR. Above the bars in i) and iii) are the FNRs and FPRs shown, respectively.

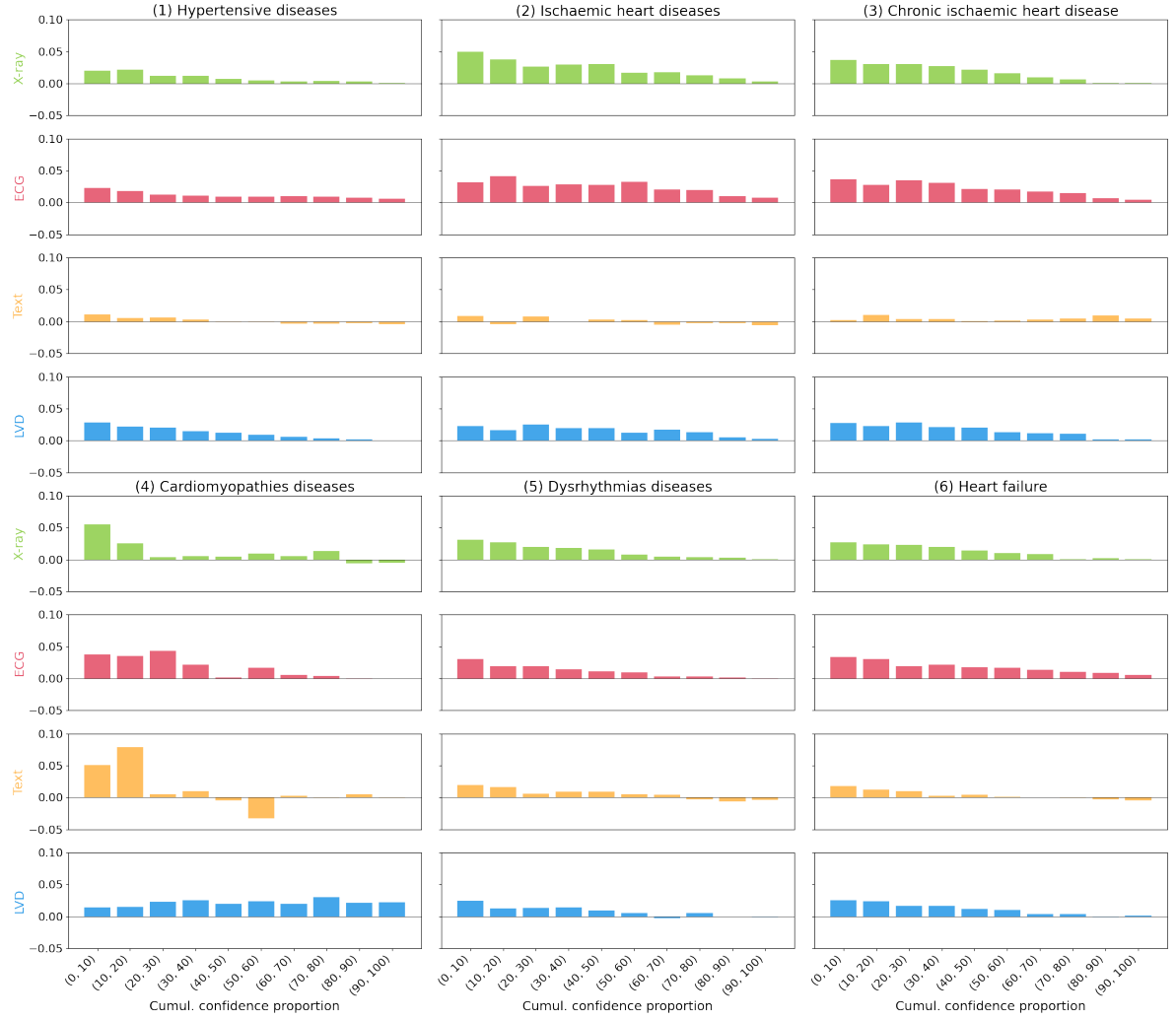


Figure 34: **Selective acquisition of multimodal data.** The figure shows the performance gain in F1-score of the multimodal model compared with the single modality models. The x-axis is divided into ten percent buckets of patients, ranked according to the confidence of the single modality model. For each bucket, the multimodal model was applied instead of the single modality model. For example, the second bucket includes the 10% of patients whose single modality predictions had the lowest 10% to 20% confidence levels. The F1-score of the multimodal model was then calculated for each bucket and compared to the F1-score achieved using only a single modality, such as X-rays. The graph indicates that, as the certainty of single-modality predictions increases (towards the right of the x-axis), the added value of the fusion model diminishes (the graph flattens out). This pattern suggests that the multimodal approach is most advantageous when single-modality predictions are less confident.

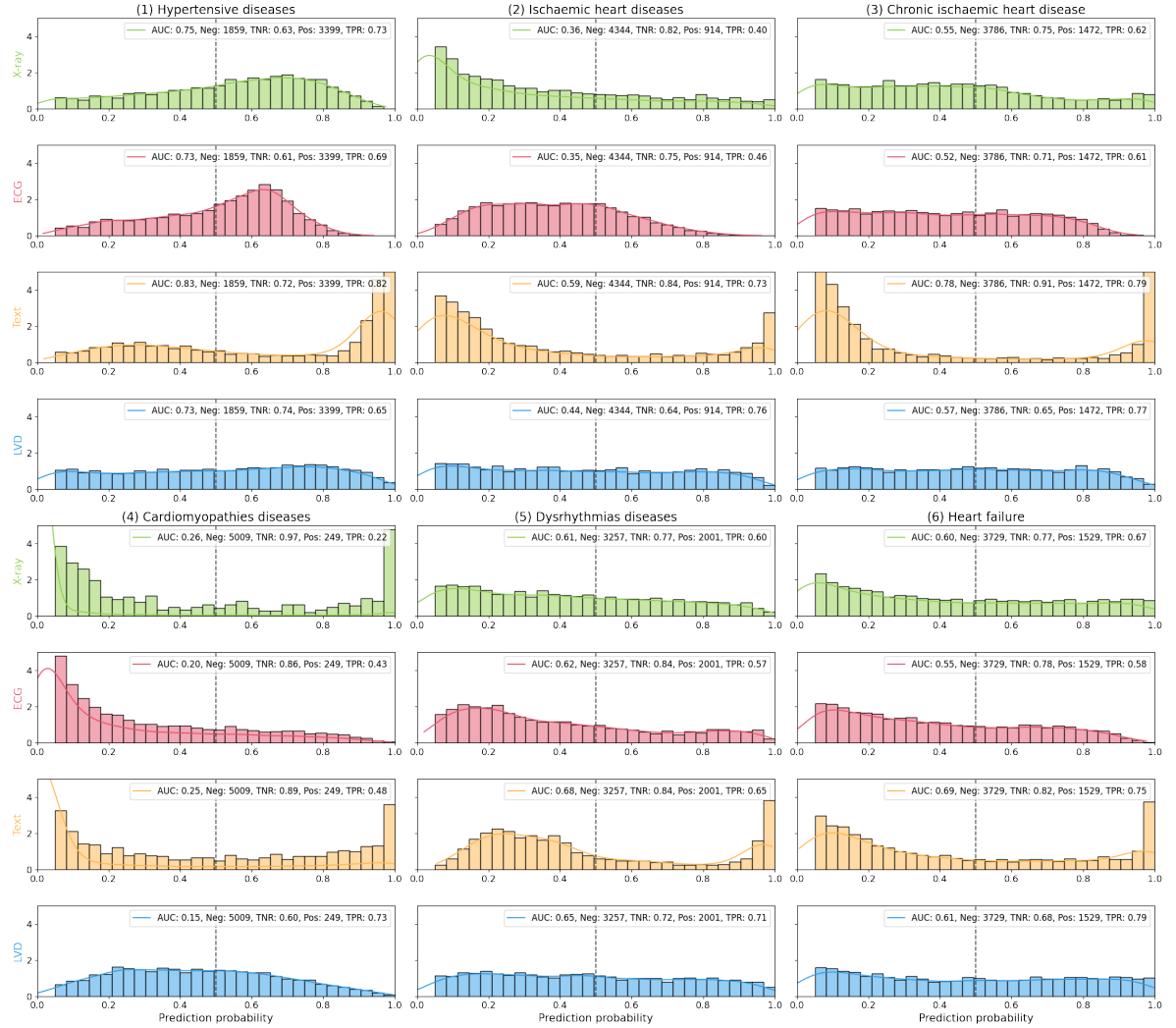


Figure 35: **Distribution of prediction probabilities.** Distribution of the prediction probabilities for the different diseases and modalities. In green for X-ray, red for ECG and blue for LVD data.

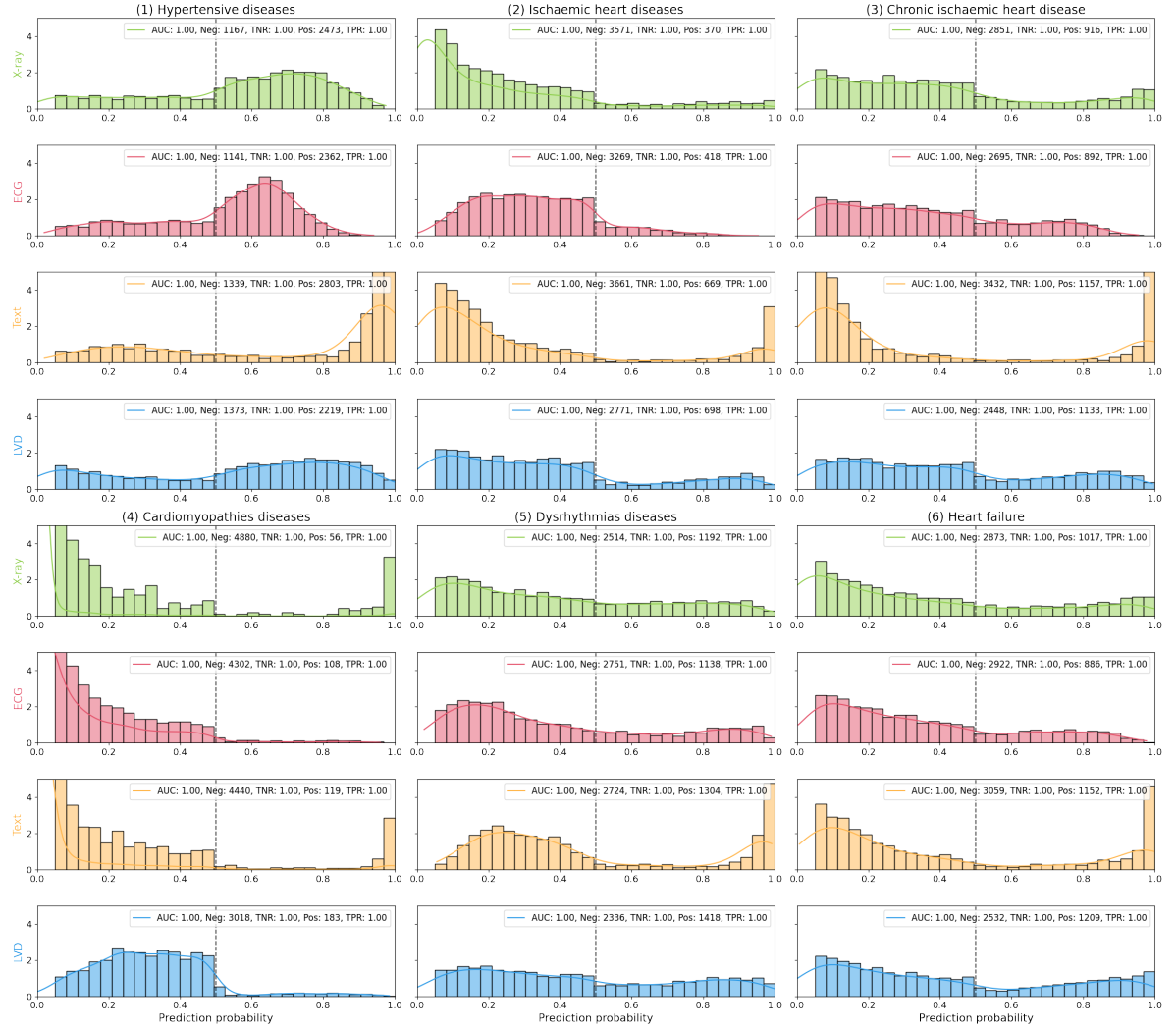


Figure 36: **Distribution of correct prediction probabilities.** Distribution of the prediction probabilities for the different diseases and modalities. In green for X-ray, red for ECG and blue for LVD data.

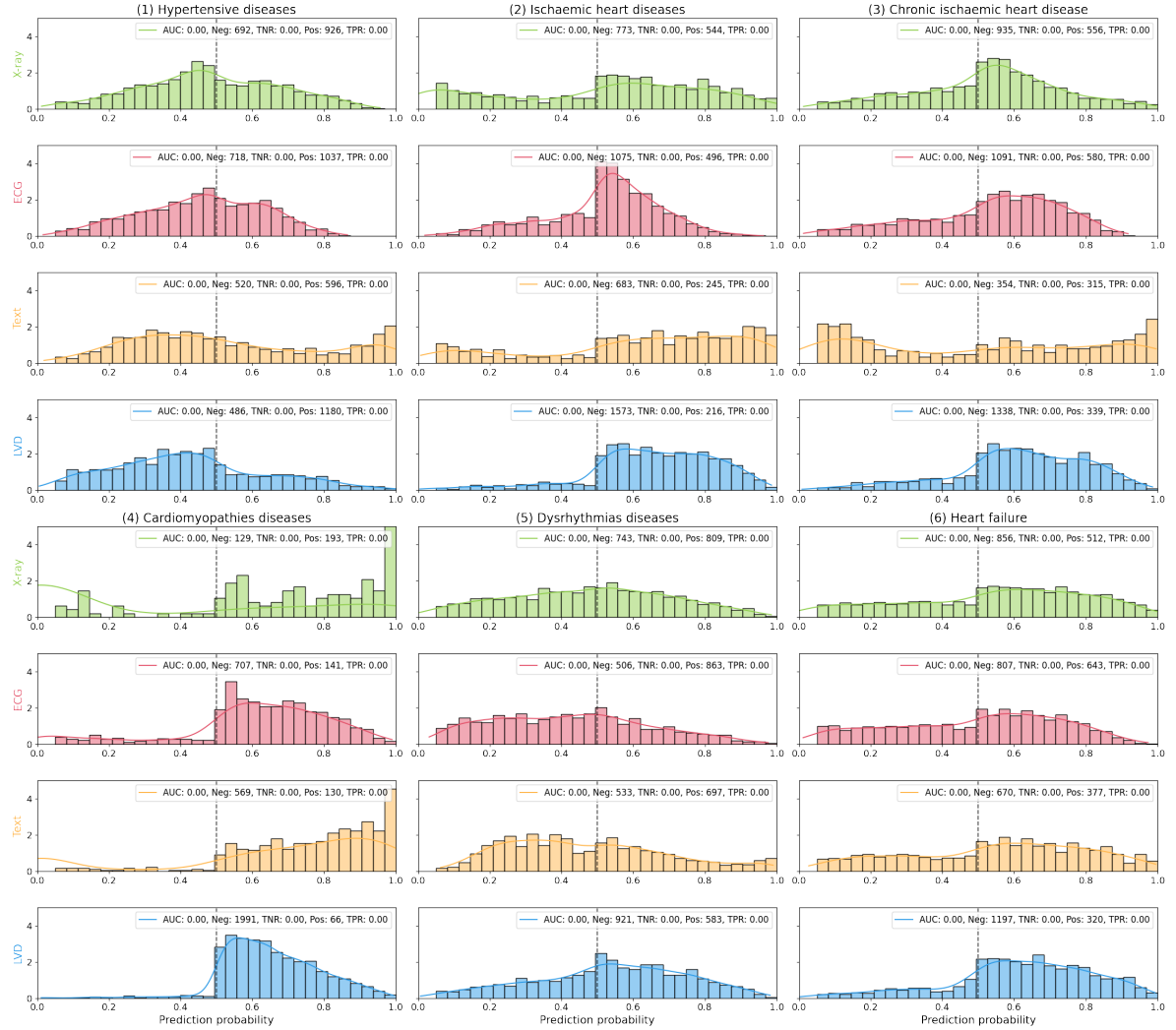


Figure 37: **Distribution of incorrect prediction probabilities.** Distribution of the prediction probabilities for the different diseases and modalities. In green for X-ray, red for ECG and blue for LVD data.

Table 32: **Per-condition confusion matrices for the best-performing model.** Counts for the late fusion model (average aggregation, decision threshold 0.5) for each of the six cardiovascular conditions (Table 11), pooled over the five cross-validation test folds ($n = 5,258$ per condition). TN, true negative, FP, false positive, FN, false negative, TP, true positive.

Condition	Actual negative		Actual positive	
	Pred. neg. (TN)	Pred. pos. (FP)	Pred. neg. (FN)	Pred. pos. (TP)
(1) Hypertensive diseases	1,314	545	904	2,495
(2) Ischaemic heart diseases	3,619	725	407	507
(3) Chronic ischaemic heart disease	2,828	958	382	1,090
(4) Cardiomyopathies diseases	4,822	187	168	81
(5) Dysrhythmias diseases	2,718	539	639	1,362
(6) Heart failure	3,014	715	411	1,118

D.3 Prompt for clinical note cleaning

The following prompt was issued to OpenAI’s gpt-4.1-mini-2025-04-14 via the OpenAI API as part of the note cleaning pipeline. The call uses structured outputs validated against a Pydantic schema, so that the model fills fixed fields rather than producing free-form text. After this prompt was applied, the *Discharge* and *Brief Hospital Course* fields were dropped, and only the remaining 11 admission-side sections were used as input to the downstream models.

You are a medical machine learning expert.

Your task is to split medical notes into sections based on when information
→ gets collected.

Do not adjust the text, just split it into sections.

Before you reply, make sure to read the entire note, understand the context
→ and that you followed the instructions below.

****SECTIONS TO SPLIT INTO:****

- Administration
- Demographics
- Allergies

- Chief complaint
- Major surgical or invasive procedure
- Medical and illness history
- Social/Family History
- Medications on admission
- Physical exams
- Imaging
- Discharge
- Brief Hospital Course:
- Other

****INSTRUCTIONS:****

1. Brief Hospital Course is everything that comes after Brief Hospital Course
 ↳ is mentioned in the note.
2. Discharge should contain all information not under Brief Hospital Course,
 ↳ but that are still discharge related, such as discharge labs, discharge
 ↳ exam, discharge medications, etc. Add all discharge related information
 ↳ to the Discharge section not just the ones I mentioned.
3. If you have exams under discharge, make sure to include them in the
 ↳ Discharge section not in the Physical exams section.
4. If you are unsure to which section a piece belongs, add it to the "Other"
 ↳ section.
5. Sections could be empty, but do not remove them.

****OUTPUT FORMAT:****

Json with two keys:

- "text": dictionary, where keys are section names and values are the text of
 ↳ the section
- "confidence": float, a confidence score between 0 and 1 indicating how
 ↳ confident you are that the text is cleaned correctly and does not contain
 ↳ any target leakage

D.4 Enabling expert conversations

MIMIC-IV and its derivatives are split across over a hundred tables, each in a different format and keyed differently (e.g. by patient, by time, by event). This fragmentation makes it not just challenging for data scientists to maintain an overview and establish meaningful connections, but it especially complicates efficient communication with domain experts, here clinicians.

To address this issue, enable future researchers to gain a quick overview and particularly facilitate discussions with experts, various tools have been developed (Figure 38): First, the problem of fragmented tables was addressed. Using structured outputs from large language models (LLMs), specifically GPT-4o (OpenAI, 2024), all dataset documentation was transformed into a standardised database markup language (DBML). This approach not only provided an easily understandable overview ((A) in Figure 38) but also revealed gaps and errors in the official MIMIC documentation by automatically detecting linkage errors in a subsequent step. In addition to soon making the created DBML available to the research community and the code to simplify future updates, the findings already led to multiple GitHub issues to resolve inaccuracies in the current MIMIC-IV documentation. Where documentation was absent, another LLM was used to generate dataset documentations from SQL structure information. Next, all used tables were standardised and merged into an event-based format ((B) in Figure 38), allowing for a clear understanding of available data over time for patients and aiding in the preparation of later analyses. Finally, a browser-based application was developed, offering a dashboard-style interface for exploring MIMIC-IV to gain rapid insights and help people without the data science experience to play with the data ((C) in Figure 38).

D.5 Data overview

Table 33: **MIMIC-IV datasets.** Overview of datasets belonging to the MIMIC-IV collection.

Dataset	Content	Type	Publication date	Version used	Version date	Collection period	Size [GB]	Patients	Observations
MIMIC-IV	Two modules: hosp: Hospital wide EHR data icu: ICU wide data	Tabular	2020/08	3.0	2024/07	2008-2022	10	364,627	
MIMIC-IV-ED	Emergency department data	Tabular	2021/06	2.2	2023/01	2011-2019	0		425,000 (stays)
MIMIC-CXR	Chest X-rays	Images	2019/08	2.1.0	2024/07	2011-2016	4700		377,110 (images)
MIMIC-CXR-JPG	Chest X-rays	Images	2019/11	2.1.0	2024/03	2011-2016	570		377,110 (images)
MIMIC-CXR-Biomarkers	Biomarkers of chest X-rays	Tabular	2024/08	1.0.0	2024/08		0		96,161 (samples)
MIMIC-IV-Notes (discharge)	Discharge summaries	Text	2023/01	2.2	2023/01		2	145,915	331,794 (reports)
MIMIC-IV-Notes (reports)	Radiology reports	Text	2023/01	2.2	2023/01		2	237,427	2,321,355 (reports)
MIMIC-IV-Ext-BHC	Standardized notes to BHC	Text	2024/09	1.0.0	2024/09		3		270,033 (notes)
MIMIC-IV-ECG	12-lead, 10sec ECGs	Waveform	2022/12	1.0	2023/09	2008-2019	90	160,000	800,000 (ECGs)
MIMIC-IV Waveform	Various signals	Waveform	2022/07	0.1.0	2022/07		13	198	200 (records)
MIMIC-IV-ECHO	Echocardiograms	Images	2023/07	0.1	July 2023	2017-2019	1700	4,579	500,000 (echod.)

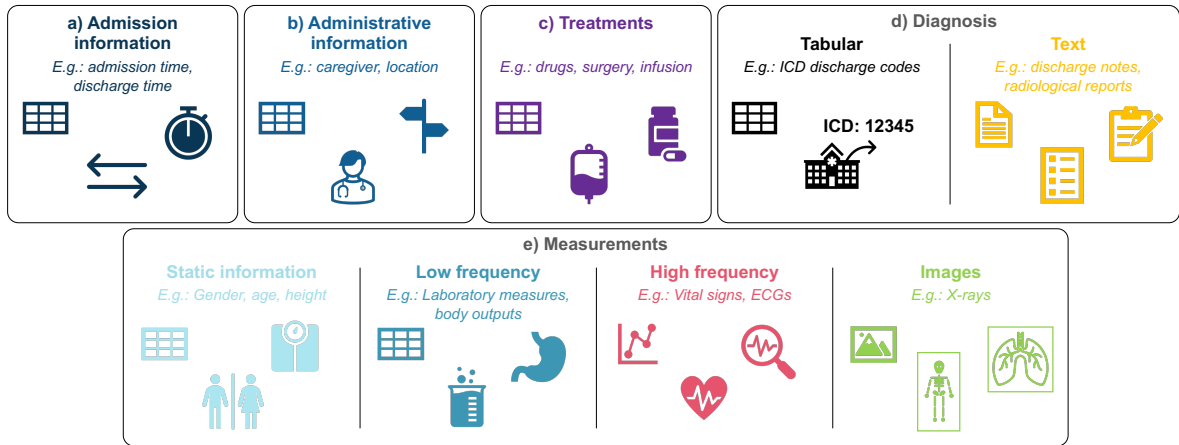


Figure 39: **Data modalities available in MIMIC-IV.** a) Admission information are available in tabular format, event-based. b) Administrative information are available for some process steps, in tabular format. c) Treatments are available as different types of information. d) Diagnosis are either available as ICD discharge codes, discharge notes or radiological reports. e) Measurements come in different forms, either as tabular data, low-frequency time-series, high-frequency time-series or images.

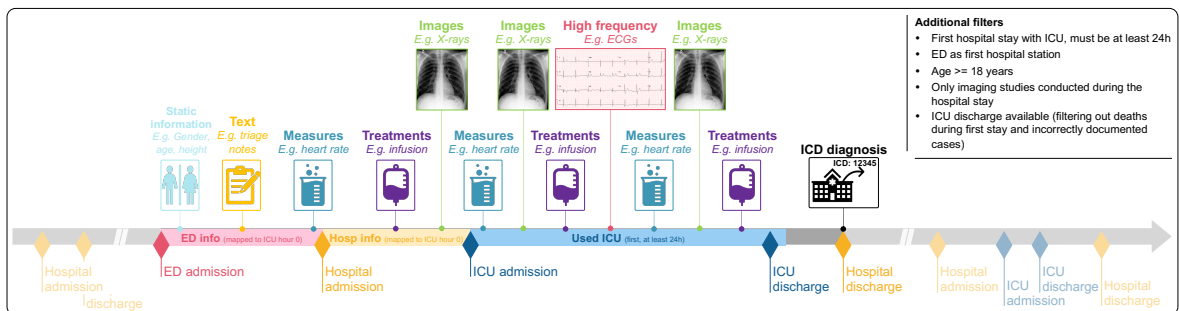


Figure 40: **Patient journey.** Most static information (such as demographics) become available with admission. High-frequency measures (i.e. ECGs) and images (i.e. X-rays) are collected at various points in time. Free text information (such as ED triage notes) are stored in MIMIC only at certain time points, similar to diagnoses, which are, for example, stored all together at the ED discharge. Additional clinical notes become available with radiological studies (i.e. radiology notes). Treatment information (such as infusions) and low-frequency measures (such as hourly heart rate or blood pressure measures) are being measured multiple times and stored aggregated. Only patients were used with an ICU stay of at least 24 hours. Only the first hospital stay of a patient that included an ICU stay was selected. Only hospital stays that started with an ED admission were considered. All data collected before the ICU stay were mapped to hour zero of the ICU stay. Only patients ≥ 18 years were selected. Only imaging studies conducted during a hospital stay were used.

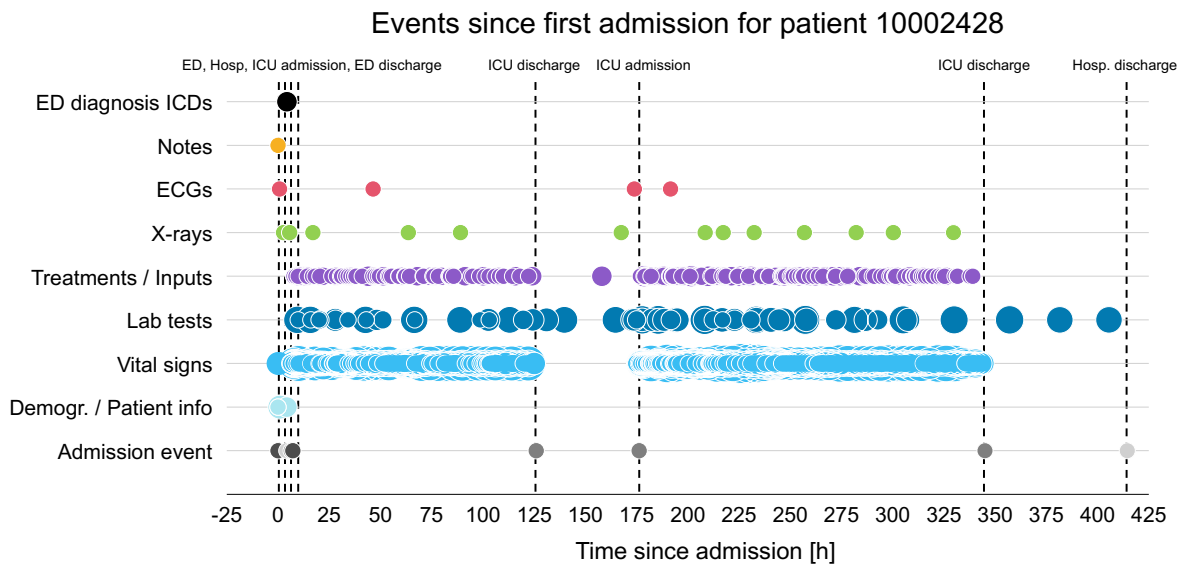


Figure 41: **Example information availability.** The figure shows an example of the time when information are tracked in MIMIC. The example is for patient 10002428. Vertical lines and the corresponding bubbles are admission and discharge events. Above are various measures. One can see that many sources of information, especially static information (e.g. gender) become available already during admission, while vital signs and lab tests are collected throughout the stay. The bubble size corresponds to the amount of information available.

E Supplementary material to Chapter 6

This section contains definitions, figures and tables with deep-dive information for Chapter 6.

Taxonomy and definitions

Shortcut learning

Spurious vs. causal relationships. I distinguish between causally relevant relationships and relationships that are invariant across environments, and I contrast these with spurious associations, which are predictive in the training environment but may not persist across deployment settings (and are typically non-causal with respect to the target).

Shortcut learning. I define shortcut learning as the phenomenon where models adopt simple, easily learned heuristics (e.g. demographic proxies or acquisition artefacts) that exploit training-set regularities and are often spurious or environment-specific, rather than relying on invariant, causally relevant features (e.g. pathognomonic patterns). Such shortcuts are typically brittle under shift.

Demographic redundancy (double counting). In multimodal fusion, I describe demographic redundancy as the state where demographic signals are provided explicitly via tabular inputs and are simultaneously encoded implicitly within the imaging modality. This redundancy can create a “double counting” effect and may amplify the influence of demographic attributes on the decision boundary.

Data characteristics and shifts

Distribution shift. Distribution shift refers to divergences between training and deployment data. I focus on two settings: *temporal shifts*, where data characteristics change over time (e.g. evolving patient mix, protocols, or devices), and *subgroup/site shifts*, where differences arise from demographics or institutions.

Fairness and robustness

Robustness vs. generalisability. I use *generalisability* to refer to the system’s ability to perform well on unseen but relevant data, such as new patients or new sites. I use *robustness* to refer to the stability of performance under specified distribution shifts or perturbations (e.g. changes in acquisition, protocols, or population composition).

Group-dependent miscalibration. This refers to systematic calibration differences where identical predicted probabilities correspond to different actual risk levels across groups (not necessarily by a constant offset).

Levelling-up vs. levelling-down. I adopt a “levelling-up” philosophy for fairness constraints ([Mittelstadt et al., 2023](#); [Zietlow et al., 2022](#)). *Levelling-down* achieves equity by reducing the performance of one or more groups, potentially including the best-off group, to reduce gaps between groups; this can arise under some group-fairness constraints or training objectives. Conversely, *levelling-up* prioritises improving (or at least not worsening) the worst-off group’s performance, and avoids avoidable degradation of other groups.

Data

Table 34: **Fusion data: Demographics.** Statistical summary of demographic characteristics in the fusion dataset. Column ‘Total’ shows the distribution of patients across the groups. The following columns show the count of positive examples per-group with the ratio of positive examples vs all in brackets. The positive count is the average positive count per-group over the different splits from cross-validation for chronic ischaemic heart disease.

Feature	Total	Pos. total	Pos. train	Pos. valid.	Pos. test	Pos. test res.
Total	5,280	1,474 (28%)	945 (28%)	234 (27%)	295 (27%)	863 (27%)
Age						
>= 63	2,870 (54%)	1,081 (38%)	690 (38%)	174 (38%)	216 (38%)	428 (27%)
< 63	2,410 (46%)	393 (16%)	255 (17%)	59 (15%)	79 (16%)	435 (27%)
Sex						
Female	2,334 (44%)	534 (23%)	345 (23%)	82 (22%)	107 (23%)	458 (29%)
Male	2,946 (56%)	940 (32%)	600 (32%)	152 (32%)	188 (32%)	449 (28%)
Insurance						
Medicaid	761 (14%)	119 (16%)	79 (16%)	16 (14%)	24 (16%)	221 (28%)
Medicare	2,944 (56%)	1,050 (36%)	672 (36%)	168 (35%)	210 (36%)	221 (28%)
Private	1,323 (25%)	273 (21%)	172 (20%)	46 (22%)	55 (21%)	224 (28%)
Other	252 (5%)	32 (13%)	22 (14%)	4 (9%)	6 (13%)	221 (28%)

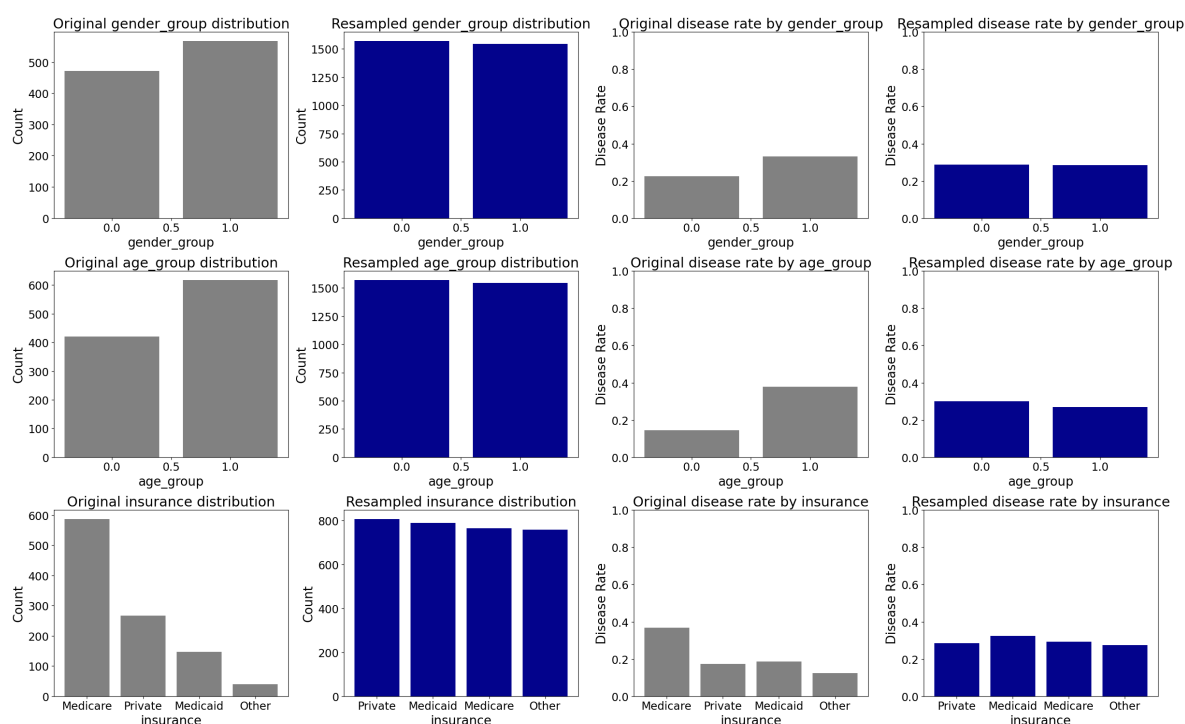


Figure 42: **Fusion data: Population characteristics of test sets.** Example characteristics for first fold. Rows: Different subgroups. First column: Original subgroup distributions. Second column: Resampled subgroup distributions. Third column: Original disease prevalence per subgroup. Fourth column: Resampled disease prevalence per subgroup. For resampling a size factor of three was used.

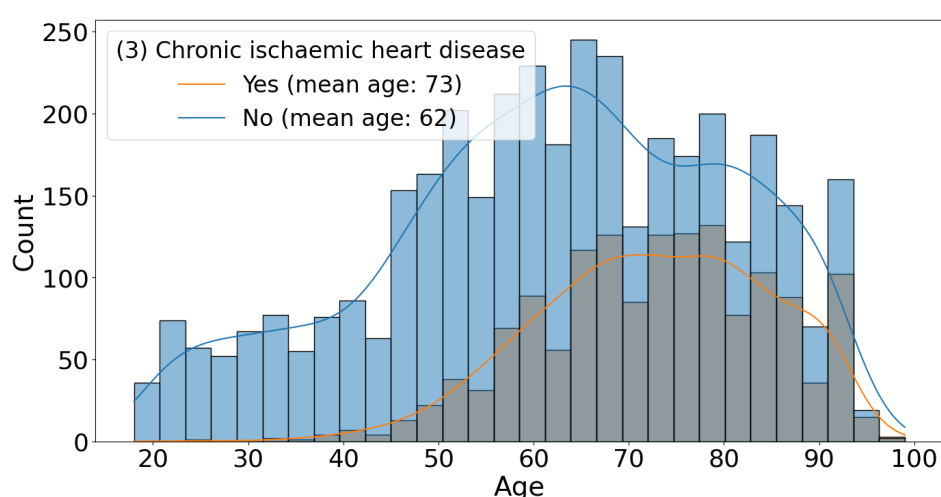


Figure 43: **Age distribution by disease group.** Example for chronic ischaemic heart disease. Patients with chronic ischaemic heart disease have a significantly higher mean age of 73 years compared to 62 years in the group without the condition.

Shortcuts and threshold calibration

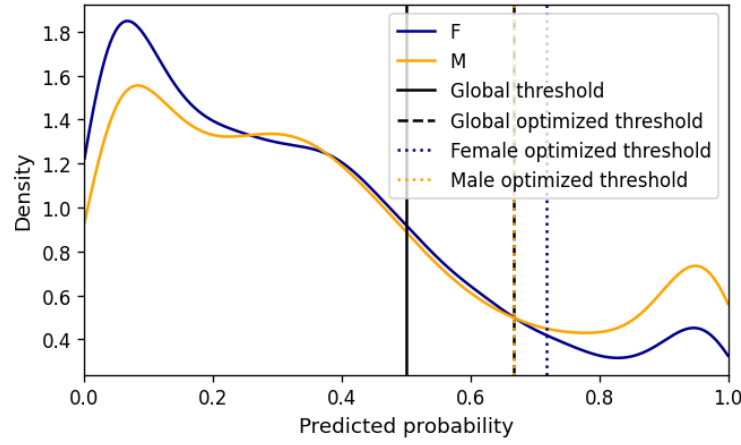


Figure 44: **Image model: Threshold shift example.** Plotted are the image model probabilities for fold 0 of the (3) Chronic ischaemic heart disease model. Only 1% of all data (2% of all female data) are affected by the group specific decision thresholds.

Table 35: **Image models: Shortcut risks.** Different model backbones are used to investigate potential model shortcuts. Chronic ischaemic heart disease is used as example disease. Multi-task models did not outperform single-task models.

Prediction task	Backbone	Pre-trained weights	AUC, mean (SD)
Disease	Single-task	ImageNet	.794 (.006)
Age	Single-task	ImageNet	.931 (.003)
Sex	Single-task	ImageNet	.999 (.000)
Disease	Multi-task	ImageNet	.794 (.008)
Age	Multi-task	ImageNet	.927 (.006)
Sex	Multi-task	ImageNet	.998 (.001)
Age	SPLIT	Random	.555 (.008)
Age	SPLIT	ImageNet	.805 (.014)
Age	SPLIT	Disease	.867 (.009)
Age	SPLIT	Disease (young only)	.828 (.017)
Sex	SPLIT	Random	.657 (.008)
Sex	SPLIT	ImageNet	.906 (.004)
Sex	SPLIT	Disease	.928 (.008)
Sex	SPLIT	Disease (male only)	.916 (.006)

Table 36: **Fusion models: Threshold dependent model performance.** Shown is the mean (SD) of accuracy (Acc), FNR and F1 for different groups of the original fusion model and the adjusted models using thresholds determined by OxonFair. Values are shown for chronic ischaemic heart disease. Darker cells indicate better values.

Group	Acc original	Acc resampled	FNR original	FNR resampled	F1 original	F1 resampled
a) Fixed decision threshold of 0.5						
Overall	.744 (.029)	N/A	.261 (.042)	N/A	.713 (.024)	N/A
Max gap	.166	.162	.111	.160	.073	.136
Worst group	.674 (.030)	.650 (.066)	.378 (.097)	.482 (.060)	.666 (.028)	.622 (.062)
Age	N/A	.721 (.042)	N/A	.384 (.039)	N/A	.673 (.030)
Age - young	.838 (.030)	.794 (.037)	.326 (.021)	.440 (.065)	.739 (.034)	.730 (.033)
Age - old	.674 (.030)	.660 (.040)	.288 (.056)	.318 (.077)	.666 (.028)	.627 (.032)
Sex	N/A	.725 (.027)	N/A	.357 (.053)	N/A	.684 (.016)
Sex - female	.735 (.018)	.714 (.026)	.337 (.059)	.433 (.065)	.674 (.018)	.665 (.023)
Sex - male	.747 (.039)	.735 (.039)	.226 (.046)	.284 (.063)	.730 (.035)	.702 (.029)
Insurance	N/A	.743 (.043)	N/A	.409 (.062)	N/A	.691 (.045)
Insurance - Medicaid	.812 (.042)	.780 (.051)	.378 (.097)	.482 (.060)	.697 (.052)	.711 (.053)
Insurance - Medicare	.690 (.031)	.650 (.066)	.277 (.054)	.322 (.026)	.679 (.027)	.622 (.062)
Insurance - Other	.856 (.068)	.812 (.074)	.306 (.108)	.381 (.179)	.739 (.087)	.758 (.096)
Insurance - Private	.802 (.042)	.778 (.022)	.293 (.063)	.355 (.082)	.733 (.034)	.732 (.021)
b) Global, for accuracy optimized decision threshold						
Overall	.788 (.012)	N/A	.540 (.084)	N/A	.703 (.027)	N/A
Max gap	.152	.056	.134	.148	.069	.086
Worst group	.716 (.017)	.740 (.019)	.663 (.261)	.705 (.225)	.731 (.037)	.710 (.019)
Age - young	.867 (.012)	.797 (.029)	.607 (.073)	.649 (.072)	.706 (.029)	.679 (.044)
Age - old	.716 (.017)	.752 (.011)	.564 (.093)	.579 (.065)	.663 (.034)	.657 (.021)
Sex - female	.794 (.015)	.740 (.019)	.624 (.070)	.688 (.058)	.662 (.016)	.622 (.033)
Sex - male	.782 (.020)	.781 (.019)	.490 (.101)	.540 (.085)	.722 (.039)	.698 (.033)
Insurance - Medicaid	.858 (.023)	.788 (.055)	.623 (.060)	.655 (.155)	.686 (.035)	.668 (.102)
Insurance - Medicare	.730 (.023)	.745 (.028)	.558 (.103)	.582 (.062)	.672 (.040)	.654 (.021)
Insurance - Other	.882 (.048)	.771 (.078)	.663 (.261)	.705 (.225)	.662 (.144)	.624 (.150)
Insurance - Private	.851 (.013)	.801 (.013)	.551 (.096)	.573 (.050)	.731 (.037)	.710 (.019)
c) Group-wise, for accuracy optimized decision threshold						
Max gap	.150	.056	.248	.226	.110	.118
Worst group	.714 (.017)	.740 (.007)	.718 (.116)	.807 (.062)	.628 (.053)	.569 (.035)
Age	.787 (.012)	.768 (.022)	.520 (.033)	.614 (.087)	.709 (.016)	.662 (.041)
Age - young	.864 (.012)	.794 (.039)	.484 (.088)	.643 (.126)	.736 (.031)	.674 (.080)
Age - old	.714 (.017)	.752 (.011)	.572 (.022)	.579 (.065)	.662 (.018)	.657 (.021)
Sex	.791 (.011)	.760 (.008)	.558 (.062)	.701 (.045)	.702 (.024)	.632 (.025)
Sex - female	.797 (.023)	.740 (.007)	.718 (.116)	.807 (.062)	.628 (.053)	.569 (.035)
Sex - male	.786 (.014)	.776 (.017)	.470 (.062)	.581 (.114)	.732 (.023)	.681 (.041)
Insurance	.787 (.012)	.772 (.023)	.507 (.061)	.646 (.086)	.711 (.023)	.659 (.044)
Insurance - Medicaid	.852 (.008)	.790 (.053)	.564 (.151)	.620 (.152)	.691 (.046)	.681 (.085)
Insurance - Medicare	.730 (.023)	.751 (.021)	.540 (.077)	.632 (.141)	.676 (.031)	.637 (.057)
Insurance - Other	.868 (.064)	.773 (.082)	.671 (.350)	.735 (.239)	.635 (.118)	.611 (.158)
Insurance - Private	.845 (.021)	.807 (.020)	.464 (.103)	.516 (.054)	.745 (.041)	.729 (.033)
d) Group-wise, for accuracy optimized decision threshold with enforced equal opportunity of 0.02						
Max gap	.142	.042	.342	.061	.090	.061
Worst group	.706 (.028)	.736 (.018)	.648 (.193)	.680 (.192)	.612 (.110)	.604 (.183)
Age - young	.853 (.028)	.794 (.032)	.430 (.026)	.604 (.061)	.737 (.030)	.692 (.041)
Age - old	.711 (.020)	.759 (.014)	.605 (.114)	.645 (.093)	.647 (.048)	.643 (.029)
Sex - female	.783 (.007)	.736 (.018)	.568 (.052)	.631 (.074)	.669 (.009)	.636 (.023)
Sex - male	.768 (.013)	.778 (.018)	.637 (.069)	.636 (.104)	.673 (.029)	.666 (.047)
Insurance - Medicaid	.810 (.054)	.765 (.066)	.433 (.305)	.646 (.241)	.663 (.080)	.633 (.147)
Insurance - Medicare	.706 (.028)	.743 (.027)	.648 (.193)	.680 (.192)	.612 (.110)	.605 (.104)
Insurance - Other	.781 (.188)	.750 (.067)	.306 (.204)	.621 (.377)	.681 (.163)	.604 (.183)
Insurance - Private	.835 (.018)	.782 (.023)	.574 (.202)	.619 (.191)	.694 (.095)	.665 (.090)
e) Group-wise, for F1 optimized decision threshold with enforced minimum recall of 0.5						
Max gap	.129	.065	.269	.185	.111	.069
Worst group	.713 (.017)	.721 (.022)	.555 (.055)	.519 (.097)	.634 (.151)	.662 (.107)
Age - young	.842 (.021)	.798 (.038)	.383 (.066)	.483 (.074)	.733 (.026)	.725 (.043)
Age - old	.713 (.017)	.733 (.027)	.555 (.055)	.485 (.076)	.664 (.023)	.663 (.023)
Sex - female	.759 (.012)	.721 (.022)	.418 (.028)	.504 (.020)	.682 (.012)	.657 (.018)
Sex - male	.785 (.013)	.776 (.011)	.470 (.070)	.479 (.107)	.731 (.022)	.706 (.032)
Insurance - Medicaid	.826 (.027)	.778 (.052)	.414 (.123)	.519 (.097)	.702 (.030)	.700 (.061)
Insurance - Medicare	.729 (.024)	.729 (.019)	.535 (.069)	.480 (.091)	.678 (.029)	.661 (.014)
Insurance - Other	.737 (.187)	.779 (.132)	.266 (.245)	.334 (.233)	.634 (.151)	.662 (.107)
Insurance - Private	.842 (.026)	.794 (.040)	.454 (.072)	.452 (.099)	.745 (.035)	.730 (.044)

Performance change over time

Methodology

To examine performance over time, we mapped de-identified dates to specific years. As MIMIC only provides mappings of anchor years to three-year intervals (2008–2010, 2011–2013, etc.) for privacy reasons, we used the middle year of each interval.

Results

Figure 45 illustrates the performance trend over time. As the models were trained on data up to and including 2012 and evaluated on subsequent years, no continuous performance degradation was observed, a risk frequently discussed in the literature (Krones and Walker, 2024). Instead, the results indicate inter-annual variability comparable to the variability observed across subgroups.

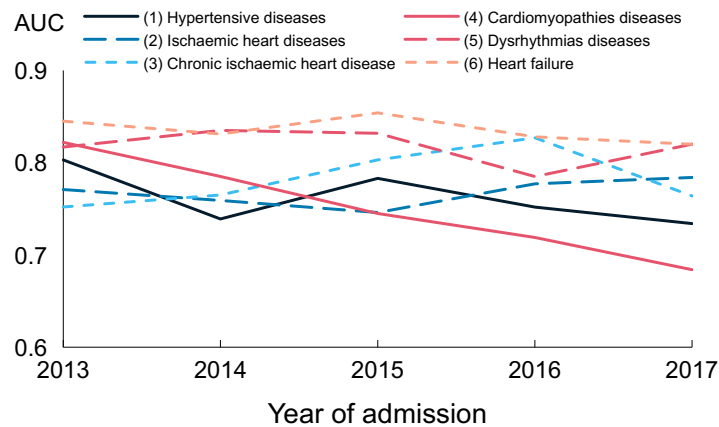


Figure 45: **Fusion models: Performance change over time.** Late fusion model trained on data with admissions between 2010 and 2012 (inclusive) and evaluated on patient data from the years thereafter. Shown is the AUC.

Adversarial mitigation

Methodology

To verify if removing shortcuts affects performance, we applied Domain-Adversarial Training (DANN) (Ganin et al., 2016). DANN trains the feature extractor to be predictive for the disease while simultaneously confusing a secondary classifier trying to predict the sensitive attribute (age or sex). If performance drops significantly under DANN, it suggests the model relies on the demographic shortcut; if it remains stable, the demographic signal is either redundant or unused.

Results

We applied DANN to encourage representations invariant to demographic attributes, training the model to perform the main predictive task while suppressing information about age or sex via an adversarial domain classifier. As reported in Table 37, this reduced the image-only model’s test AUC when the adversary targeted sex ($0.794 \rightarrow 0.788$), whereas the age-targeted image model and both fusion models changed only slightly ($0.824 \rightarrow 0.822$). These findings are consistent with the image model exploiting sex-correlated signals: DANN pushes the representation towards sex invariance and thus reduces reliance on those signals, whereas the fusion model was less affected, likely because demographic information is redundantly available across modalities. Because the differences are small relative to cross-fold variability (see SDs in Table 37) and performance on the test distribution did not increase, we did not observe improved robustness in terms of AUC. Finally, adversarial learning should be interpreted with caution: as shown in Figure 46, training outcomes are strongly dependent on the domain-regulation weight (λ).

Table 37: **Model comparison: Disease prediction performance after DANN training.** Performance of the image, ECG and fusion models after applying Domain-Adversarial Training of Neural Networks (DANN) to suppress demographic information in the learned representations. ‘Domain feature’ denotes the attribute (e.g. sex or age) whose influence was mitigated using DANN. Shown is the mean (SD) AUC.

Model	Domain feature	Chronic ischaemic heart disease	Hypertensive disease
Image	None	.794 (.006)	-
Fusion	None	.824 (.016)	-
ECG	None	-	.709 (.010)
Image	Age	.793 (.006)	-
Fusion	Age	.822 (.013)	-
ECG	Age	-	.708 (.008)
Image	Sex	.788 (.004)	-
Fusion	Sex	.822 (.012)	-

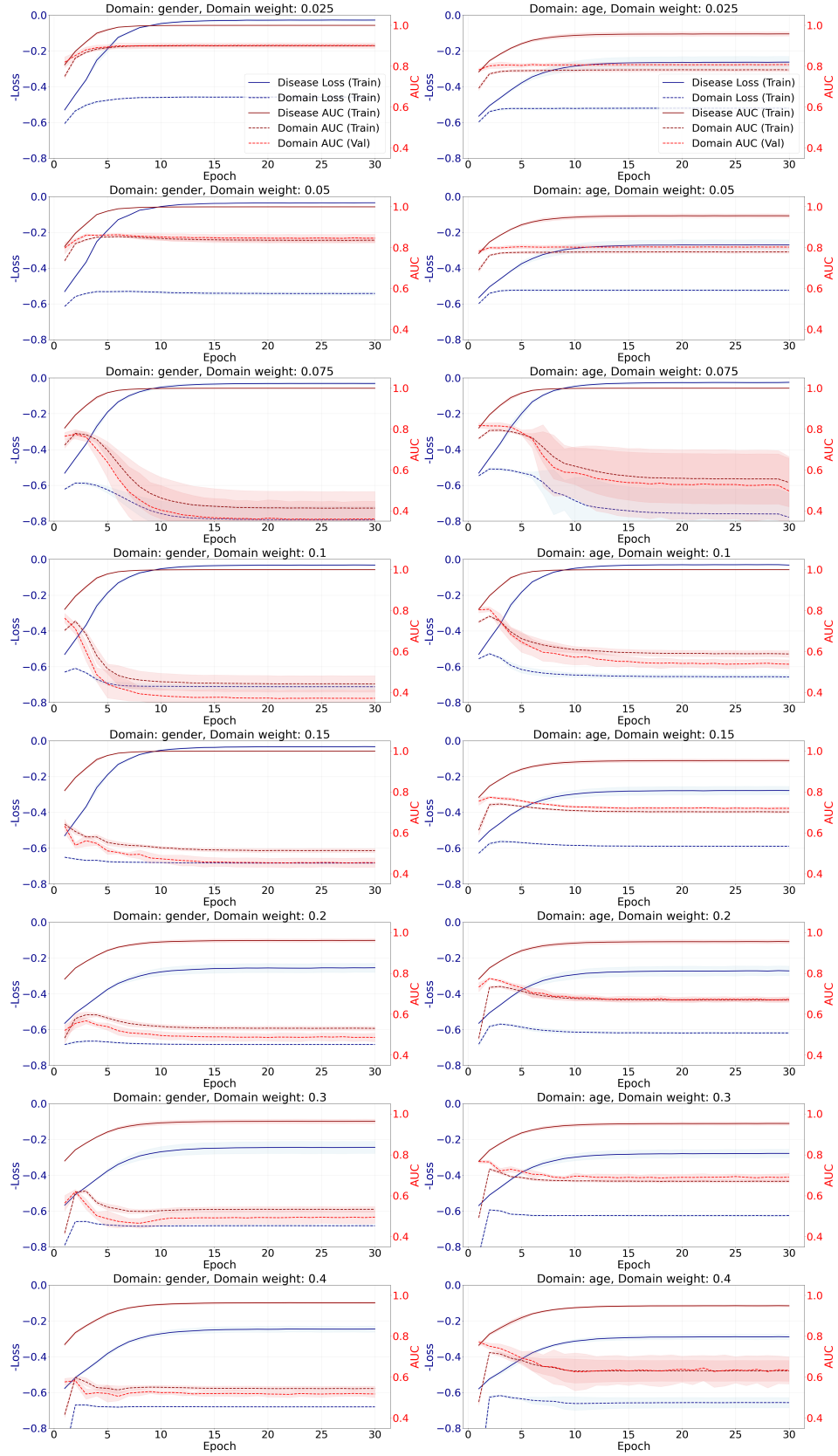


Figure 46: **Image model: DANN performance by regulation weight.** Plotted is the negative loss and AUC for disease (chronic ischaemic heart disease) and domain for different weights of the regulation term across training epochs. Plotted are the average (and filled standard deviation) over the different folds.

References

- M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.
- R. Achalia, A. Sinha, A. Jacob, G. Achalia, V. Kaginalkar, G. Venkatasubramanian, and N. P. Rao. A proof of concept machine learning analysis using multimodal neuroimaging and neurocognitive measures as predictive biomarker in bipolar disorder. *Asian Journal of Psychiatry*, 50:101984, 2020.
- J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.
- A. Adam, A. K. Dixon, J. H. Gillard, and C. Schaefer-Prokop. *Grainger & Allison’s diagnostic radiology, 2 volume set e-book*. Elsevier Health Sciences, 2020.
- R. Adams, K. E. Henry, A. Sridharan, H. Soleimani, A. Zhan, N. Rawat, L. Johnson, D. N. Hager, S. E. Cosgrove, A. Markowski, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine*, 28(7):1455–1460, 2022.
- J. Adleberg, A. Wardeh, F. X. Doo, B. Marinelli, T. S. Cook, D. S. Mendelson, and A. Kagen. Predicting patient demographics from chest radiographs with deep learning. *Journal of the American College of Radiology*, 19(10):1151–1161, 2022.
- A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.
- M. M. Ahsan and Z. Siddique. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 128:102289, 2022.
- A. Akay and H. Hess. Deep learning: Current and emerging applications in medicine and technology. *IEEE Journal of Biomedical and Health Informatics*, 23(3):906–920, 2019.
- H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.
- N. Alam, E. L. Hobbelink, A. J. van Tienhoven, P. M. van de Ven, E. P. Jansma, and P. W. B. Nanayakkara. The impact of the use of the Early Warning Score (EWS) on patient outcomes: A systematic review. *Resuscitation*, 85(5):587–594, 2014.

- J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.
- J. M. L. Alcaraz, H. Bouma, and N. Strodthoff. Enhancing clinical decision support with physiological waveforms—a multimodal benchmark in emergency care. *Computers in Biology and Medicine*, 192: 110196, 2025.
- A. Alloula, R. Mustafa, D. R. McGowan, and B. W. Papież. On biases in a UK Biobank-based retinal image classification model. *arXiv:2408.02676*, 2024.
- A. Alloula, C. Jones, B. Glocker, and B. W. Papież. Subgroups matter for robust bias mitigation. *arXiv:2505.21363*, 2025.
- S. Amal, L. Safarnejad, J. A. Omiye, I. Ghanzouri, J. H. Cabot, and E. G. Ross. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine*, 9, 2022.
- E. Amorim, W.-L. Zheng, M. M. Ghassemi, M. Aghaeeval, P. Kandhare, V. Karukonda, J. W. Lee, S. T. Herman, A. Sivaraju, N. Gaspard, J. Hofmeijer, M. J. A. M. van Putten, R. Sameni, M. A. Reyna, G. D. Clifford, and M. B. Westover. The international cardiac arrest research (I-CARE) consortium electroencephalography database. In *Critical Care Medicine 2023*, page (in press), 2023.
- S. Anandaciva. Critical care services in the English NHS — kingsfund.org.uk. <https://www.kingsfund.org.uk/insight-and-analysis/long-reads/critical-care-services-nhs>, 2020. Accessed in Oct 2024.
- S. Ö. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.
- L. C. Armitage, S. Davidson, A. Mahdi, M. Harford, R. McManus, A. Farmer, P. Watkinson, and L. Tarassenko. Diagnosing hypertension in primary care: A retrospective cohort study to investigate the importance of night-time blood pressure assessment. *British Journal of General Practice*, 73 (726):e16–e23, 2023.
- S. Atito, M. Awais, and J. Kittler. Sit: Self-supervised vision transformer. *arXiv:2104.03602*, 2021.
- S. Atito, M. Awais, and J. Kittler. GMML is all you need. *arXiv:2205.14986*, 2022.
- S. Ayesha, M. K. Hanif, and R. Talib. Performance enhancement of predictive analytics for health informatics using dimensionality reduction techniques and fusion frameworks. *IEEE Access*, 2021.
- M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry, and

- A. H. Gandomi. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in Biology and Medicine*, page 105253, 2022.
- S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.
- S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, P. MacWilliams, S. S. Mahdavi, E. Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv:2205.09723*, 2022.
- F. Badilini, T. Erdem, W. Zareba, and A. J. Moss. ECGScan: A method for conversion of paper electrocardiographic printouts to digital electrocardiographic files. *Journal of electrocardiology*, 38(4):310–318, 2005.
- A. Bagheri, T. K. J. Groenhof, W. B. Veldhuis, P. A. de Jong, F. W. Asselbergs, and D. L. Oberski. Multimodal learning for cardiovascular risk prediction using EHR data. *arXiv:2008.11979*, 2020.
- R. Balakrishnan and R. Priya. Multimodal medical image fusion based on deep learning neural network for clinical treatment analysis. *International Journal of ChemTech Research*, 11(06):160–176, 2018.
- T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.
- I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach. Comparison of deep learning approaches for multi-label chest X-Ray classification. *Scientific Reports*, 9(1):1–10, 2019.
- I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, et al. Reading Race: AI recognises patient’s racial identity in medical images. *arXiv:2107.10356*, 2021.
- J. Barnard and X.-L. Meng. Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1):17–36, 1999.
- I. Bartoletti. AI in healthcare: Ethical and privacy challenges. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 7–10. Springer, 2019.
- A. Barua, M. U. Ahmed, and S. Begum. A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions. *IEEE Access*, 11:14804–14831, 2023.
- S. Basu, S. Singhal, and D. Singh. A systematic literature review on multimodal medical image fusion. *Multimedia Tools and Applications*, 83(6):15845–15913, 2024.

- J. J. Battista, W. D. Rider, and J. Van Dyk. Computed tomography for radiotherapy planning. *International Journal of Radiation Oncology, Biology, Physics*, 6(1):99–107, 1980.
- M. Baydoun, L. Safatly, O. K. Abou Hassan, H. Ghaziri, A. El Hajj, and H. Isma’eel. High precision digitization of paper-based ECG records: A step toward machine learning. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–8, 2019.
- K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021.
- A. M. Bean, K. Korgul, F. Krones, R. McCraith, and A. Mahdi. Exploring the landscape of large language models in medical question answering. *arXiv:2310.07225*, 2024.
- E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.
- F. Behrad and M. S. Abadeh. An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications*, page 117006, 2022.
- B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermesen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.
- X. Bi, S. Tang, Z. Yang, X. Deng, B. Xiao, and P. Liò. MMCTNet: Multi-modal Cony-Transformer network for predicting good and poor outcomes in cardiac arrest patients. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- W. D. Bidgood Jr, S. C. Horii, F. W. Prior, and D. E. Van Syckle. Understanding and using DICOM, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, 4(3):199–212, 1997.
- P. Bizopoulos and D. Koutsouris. Deep learning in cardiology. *IEEE Reviews in Biomedical Engineering*, 12:168–193, 2018.
- O. Bohdal, T. Hospedales, P. H. Torr, and F. Barez. Fairness in AI and its long-term implications on society. *arXiv:2304.09826*, 2023.
- R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg,

- A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.
- N. Braman, J. W. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe, and J. Venkataraman. Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 667–677. Springer, 2021.
- W. E. Brant and C. A. Helms. *Fundamentals of diagnostic radiology*. Lippincott Williams & Wilkins, 2012.
- L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir. Machine learning for medical ultrasound: Status, methods, and future opportunities. *Abdominal Radiology*, 43:786–799, 2018.
- D. Bremner. *Brain imaging handbook*. WW Norton & Co, 2005.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- G. Brugnara, U. Neuberger, M. A. Mahmutoglu, M. Foltyn, C. Herweh, S. Nagel, S. Schönenberger, S. Heiland, C. Ulfert, P. A. Ringleb, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke*, 51(12):3541–3551, 2020.
- A. Brunelli, A. Charloux, C. T. Bolliger, G. Rocco, J.-P. Sculier, G. Varela, M. Licker, M. Ferguson, C. Faivre-Finn, R. M. Huber, et al. ERS/ESTS clinical guidelines on fitness for radical therapy in lung cancer patients (surgery and chemo-radiotherapy). *European Respiratory Journal*, 34(1):17–41, 2009.
- J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.
- A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá. Padchest: A large chest X-Ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.
- T. M. Buzug. Computed tomography. In *Springer Handbook of Medical Technology*, pages 311–342. Springer, 2011.
- N. Cahan, E. Klang, E. M. Marom, S. Soffer, Y. Barash, E. Burshtein, E. Konen, and H. Greenspan. Multimodal fusion models for pulmonary embolism mortality prediction. *Scientific Reports*, 13(1):7544, 2023.

- E. Çalh, 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.
- J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 6:679–698, 1986.
- M. Cao, M. Yang, C. Qin, X. Zhu, Y. Chen, J. Wang, and T. Liu. Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data. *Biomedical Signal Processing and Control*, 70:103015, 2021.
- A. Carovac, F. Smajlovic, and D. Junuzovic. Application of ultrasound in medicine. *Acta Informatica Medica*, 19:168 – 171, 2011. URL <https://api.semanticscholar.org/CorpusID:14705887>.
- A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, et al. A systematic review of natural language processing applied to radiology reports. *BMC medical informatics and decision making*, 21(1):179, 2021.
- F. Ceccarelli and M. Mahmoud. Multimodal temporal machine learning for bipolar disorder and depression recognition. *Pattern Analysis and Applications*, 25(3):493–504, 2022.
- I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. *arXiv:2010.02559*, 2020.
- G. Chauhan, R. Liao, W. Wells, J. Andreas, X. Wang, S. Berkowitz, S. Horng, P. Szolovits, and P. Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2020.
- Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.
- A. Cheerla and O. Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019.
- F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu. VLP: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.
- R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 2020.
- R. J. Chen, T. Y. Chen, J. Lipkova, J. J. Wang, D. F. Williamson, M. Y. Lu, S. Sahai, and F. Mahmood. Algorithm fairness in AI for medicine and healthcare. *arXiv:2110.00603*, 2021a.

- R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4025, October 2021b.
- R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.
- Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18, 2015.
- A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.
- M. Chua, D. Kim, J. Choi, N. G. Lee, V. Deshpande, J. Schwab, M. H. Lev, R. G. Gonzalez, M. S. Gee, and S. Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.
- G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology challenge 2016. In *2016 Computing in Cardiology (CinC)*, pages 609–612, 2016.
- G. S. Collins, K. G. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. Van Smeden, et al. TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385, 2024.
- D. J. Cook, S. Webb, and A. Proudfoot. Assessment and management of cardiovascular disease in the intensive care unit. *Heart*, 108(5):397–405, 2022.
- C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review. *arXiv:2203.15588*, 2022.
- M. Czosnyka and J. D. Pickard. Monitoring and interpretation of intracranial pressure. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(6):813–821, 2004.
- M. De-Arteaga, R. Fogliato, and A. Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

- J. Deasy, P. Liò, and A. Ercole. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. *Scientific Reports*, 10(1):22129, 2020.
- Deepanshi, K. Shivashankara, G. D. Clifford, M. A. Reyna, and R. Sameni. ECG-Image-Kit: A toolkit for synthesis, analysis, and digitization of electrocardiogram images, January 2024.
- E. Delaney, Z. Fu, S. Wachter, B. Mittelstadt, and C. Russell. OxonFair: A flexible toolkit for algorithmic fairness. *arXiv:2407.13710*, 2024.
- A. Demolder, V. Kresnakova, M. Hojcka, V. Boza, A. Iring, A. Rafajdus, S. Rovder, T. Palus, M. Herman, F. Bauer, et al. High precision ECG digitization using artificial intelligence. *Journal of Electrocardiology*, 90:153900, 2025.
- A. Dempster, F. Petitjean, and G. I. Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34: 1454–1495, 2020.
- P. P. Dendy and B. Heaton. *Physics for diagnostic radiology*. CRC Press, 2011.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.
- A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014.
- F. M. Dias, E. Ribeiro, Q. B. Soares, J. E. Krieger, and M. A. Gutierrez. Image-based electrocardiogram classification using pre-trained ConvNext. In *2024 Computing in Cardiology (CinC)*, 2024.
- H. Doğan, A. de Roos, J. Geleijns, M. V. Huisman, and L. J. Kroft. The role of computed tomography in the diagnosis of acute and chronic pulmonary embolism. *Diagnostic and Interventional Radiology*, 21(4):307, 2015.
- J. M. Dolly and A. Nisa. A survey on different multimodal medical image fusion techniques and methods. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, pages 1–5. IEEE, 2019.
- A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani,

- M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.
- K. L. Dragan, S. M. Desai, J. Billings, and S. A. Glied. Association of insurance mix and diagnostic coding practices in New York State hospitals. In *JAMA Health Forum*, volume 3, pages e222919–e222919. American Medical Association, 2022.
- H. Duanmu, P. B. Huang, S. Brahmavar, S. Lin, T. Ren, J. Kong, F. Wang, and T. Q. Duong. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 242–252. Springer, 2020.
- V. Duclos, A. Iep, L. Gomez, L. Goldfarb, and F. L. Besson. PET molecular imaging: A holistic review of current practice and emerging perspectives for diagnosis, therapeutic evaluation and prognosis in clinical oncology. *International Journal of Molecular Sciences*, 22(8):4159, 2021.
- R. O. Duda and P. E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.
- B. Duvieusart, F. Krones, G. Parsons, L. Tarassenko, B. W. Papież, and A. Mahdi. Multimodal cardiomegaly classification with image-derived digital biomarkers. In *Annual Conference on Medical Image Understanding and Analysis*, pages 13–27. Springer, 2022.
- B. Duvieusart, F. Krones, G. Parsons, L. Tarassenko, B. W. Papież, and A. Mahdi. Heart and lung segmentations for MIMIC-CXR/MIMIC-CXR-JPG and Montgomery County TB databases. *PhysioNet*, 2023.
- B. Duvieusart, F. Krones, G. Parsons, L. Tarassenko, B. W. Papież, and A. Mahdi. Image-derived cardiomegaly biomarker values for 96K chest X-rays in MIMIC-CXR/MIMIC-CXR-JPG. *PhysioNet*, 2024.
- S. Dwivedi, T. Goel, M. Tanveer, R. Murugan, and R. Sharma. Multi-modal fusion based deep learning network for effective diagnosis of Alzheimers disease. *IEEE MultiMedia*, 2022.
- S. El-Sappagh, T. Abuhmed, S. R. Islam, and K. S. Kwak. Multimodal multitask deep learning model for Alzheimer’s disease progression detection based on time series data. *Neurocomputing*, 412:197–215, 2020.
- L. Ericsson, H. Gouk, and T. M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5414–5423, June 2021.

- L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.
- G. J. Escobar, V. X. Liu, A. Schuler, B. Lawson, J. D. Greene, and P. Kipnis. Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20):1951–1960, 2020.
- A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.
- C. Fang, S. Bai, Q. Chen, Y. Zhou, L. Xia, L. Qin, S. Gong, X. Xie, C. Zhou, D. Tu, et al. Deep learning for predicting COVID-19 malignant progression. *Medical Image Analysis*, 72:102096, 2021.
- M. Fatima, M. Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.
- FDA. Artificial intelligence-enabled medical devices — fda.gov. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>, 2025. Accessed in Feb 2026.
- N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):1–13, 2022.
- P. B. Forgacs, O. Devinsky, and N. D. Schiff. Independent functional outcomes after prolonged coma following cardiac arrest: A mechanistic hypothesis. *Annals of Neurology*, 87(4):618–632, 2020.
- J. D. Fortune, N. E. Coppa, K. T. Haq, H. Patel, and L. G. Tereshchenko. Digitizing ECG image: New fully automated method and open-source software code. *medRxiv*, pages 2021–07, 2021.
- FotoFinder Systems GmbH. FotoFinder ATBM master – KI-gestütztes Total Body Mapping zur Hautkrebsfrüherkennung — fotofinder.de. <https://www.fotofinder.de/produkte/atbm-master>, 2026. Accessed in Feb 2026.
- R. K. Freeman, T. J. Garite, M. P. Nageotte, and L. A. Miller. *Fetal heart rate monitoring*. Lippincott Williams & Wilkins, 2012.
- G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.
- Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

- J. Gao, P. Li, Z. Chen, and J. Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.
- J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pages 1–77, 2023.
- N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer. Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, 7:100864, 2020.
- S. Ghosal, Q. Chen, G. Pergola, A. L. Goldman, W. Ulrich, K. F. Berman, G. Blasi, L. Fazio, A. Rampino, A. Bertolino, et al. G-MIND: An end-to-end multimodal imaging-genetics framework for biomarker identification and disease classification. In *Medical Imaging 2021: Image Processing*, volume 11596, page 115960C. International Society for Optics and Photonics, 2021.
- B. Glocker, C. Jones, M. Bernhardt, and S. Winzeck. Algorithmic encoding of protected characteristics in image-based models for disease detection. *arXiv:2110.14755*, 2021.
- B. Glocker, C. Jones, M. Bernhardt, and S. Winzeck. Risk of bias in chest X-ray foundation models. *arXiv:2209.02965*, 2022.
- N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2024.
- A. Golden, S. Hsia, F. Sun, B. Acun, B. Hosmer, Y. Lee, Z. DeVito, J. Johnson, G.-Y. Wei, D. Brooks, et al. Generative AI beyond LLMs: System implications of multi-modal generation. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 257–267. IEEE, 2024.
- L. W. Goldman. Principles of CT and CT technology. *Journal of Nuclear Medicine Technology*, 35(3): 115–128, 2007.
- F. A. Gonzalez, C. Santonocito, T. Lamas, P. Costa, S. M. Vieira, H. A. Ferreira, and F. Sanfilippo. Is artificial intelligence prepared for the 24-h shifts in the ICU? *Anaesthesia Critical Care & Pain Medicine*, 43(6):101431, 2024.
- I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.

- P. C. Gøtzsche and K. J. Jørgensen. Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*, 2013.
- B. Gow, T. Pollard, L. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J. Waks, P. Eslami, T. Carbonati, et al. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset (version 1.0). *PhysioNet*, 2023.
- B. Gow, T. Pollard, N. Greenbaum, B. Moody, A. Han, J. W. Waks, A. Johnson, E. Herbst, P. Eslami, A. Chaudhari, T. Carbonati, S. Berkowitz, R. Mark, and S. Horng. MIMIC-IV-ECHO: Echocardiogram matched subset. *PhysioNet*, 2026. Version 1.0.
- D. Grant, B. W. Papież, G. Parsons, L. Tarassenko, and A. Mahdi. Deep learning classification of cardiomegaly using combined imaging and non-imaging ICU data. In *Annual Conference on Medical Image Understanding and Analysis*, pages 547–558. Springer, 2021.
- T. Greenhalgh, J. Wherton, C. Papoutsis, J. Lynch, G. Hughes, S. Hinder, N. Fahy, R. Procter, S. Shaw, et al. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet research*, 19(11):e8775, 2017.
- V. P. Grover, J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. McPhail. Magnetic resonance imaging: Principles and techniques: Lessons for clinicians. *Journal of Clinical and Experimental Hepatology*, 5(3):246–255, 2015.
- H. Gu, H. Dong, J. Yang, and M. A. Mazurowski. How to build the best medical image segmentation algorithm using foundation models: A comprehensive empirical study with Segment Anything Model. *arXiv:2404.09957*, 2024.
- A. Guermazi, F. W. Roemer, I. K. Haugen, M. D. Crema, and D. Hayashi. MRI-based semiquantitative scoring of joint pathology in osteoarthritis. *Nature Reviews Rheumatology*, 9(4):236–251, 2013.
- Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019.
- F. Haghighi, M. R. H. Taher, M. B. Gotway, and J. Liang. DiRA: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022.
- R. Han, J. N. Acosta, Z. Shakeri, J. Ioannidis, E. Topol, and P. Rajpurkar. Randomized controlled trials evaluating AI in clinical practice: A scoping evaluation. *medRxiv*, pages 2023–09, 2023.
- E. Handley. £2 billion NHS digitisation implementation delayed by a year — openaccessgovernment.org.

- <https://www.openaccessgovernment.org/nhs-digitisation-implementation-delayed/164584/>, 2023. Accessed in Aug 2023.
- R. Hashemi, W. Bradley, and C. Lisanti. Basic principles of MRI. In *MRI: The Basics*, pages 16–31. Lippincott Williams & Wilkins, Philadelphia, 2010.
- T. Hastie, R. Tibshirani, and J. Friedman. Overview of supervised learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pages 9–41, 2009.
- N. Hayat, K. J. Geras, and F. E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR, 2022.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.
- S. He, Y. Nie, H. Wang, S. Yang, Y. Wang, Z. Cai, Z. Chen, Y. Xu, L. Luo, H. Xiang, X. Lin, M. Wu, Y. Peng, G. Shih, Z. Xu, X. Wu, Q. Wang, R. C. K. Chan, V. Vardhanabhuti, W. C. W. Chu, Y. Zheng, P. Rajpurkar, K. Zhang, and H. Chen. Towards generalizable AI in medicine via Generalist–Specialist Collaboration. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01653-3.
- Healthvisors. IDx-DR – Autonomous AI System for Diabetic Retinopathy Detection — healthvisors.com. <https://www.healthvisors.com/en/idx-dr/>, 2026. Accessed in Feb 2026.
- L. Heiliger, A. Sekuboyina, B. Menze, J. Egger, and J. Kleesiek. Beyond medical imaging—a review of multimodal deep learning in radiology. *TechRxiv*, 2022.
- P. Hempel, P. Zschke, M. Goldammer, and N. Spicher. Fusion of features with neural networks for prediction of secondary neurological outcome after cardiac arrest. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019a.
- D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing systems*, 32, 2019b.
- M. Herdman, C. Gudex, A. Lloyd, M. Janssen, P. Kind, D. Parkin, G. Bonsel, and X. Badia. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, 20:1727–1736, 2011.
- H. Hermessi, O. Mourali, and E. Zagrouba. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183:108036, 2021.

- A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.
- L. J. Hirsch, M. W. Fong, M. Leitinger, S. M. LaRoche, S. Beniczky, N. S. Abend, J. W. Lee, C. J. Wusthoff, C. D. Hahn, M. B. Westover, et al. American clinical neurophysiology society’s standardized critical care EEG terminology: 2021 Version. *Journal of clinical neurophysiology*, 38(1):1–29, 2021.
- J. Hofmeijer, T. M. Beernink, F. H. Bosch, A. Beishuizen, M. C. Tjepkema-Cloostermans, and M. J. van Putten. Early EEG contributes to multimodal outcome prediction of postanoxic coma. *Neurology*, 85(2):137–143, 2015.
- G. Holste, S. C. Partridge, H. Rahbar, D. Biswas, C. I. Lee, and A. M. Alessio. End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3294–3303, 2021.
- M. R. Hosseinzadeh Taher, F. Haghighi, R. Feng, M. B. Gotway, and J. Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021*, pages 3–13. Springer, 2021.
- P. V. Hough. Method and means for recognizing complex patterns, Dec. 18 1962. US Patent 3,069,654.
- A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.
- B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong. A review of multimodal medical image fusion techniques. *Computational and Mathematical Methods in Medicine*, 2020, 2020a.
- G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.
- S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digital Medicine*, 3(1):1–9, 2020b.
- S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: A case-study in pulmonary embolism detection. *Scientific Reports*, 10(1):1–9, 2020c.
- S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. GLoRIA: A multimodal global-local representation

- learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.
- S.-L. Huang, X. Xu, and L. Zheng. An information-theoretic approach to unsupervised feature selection for high-dimensional data. *IEEE Journal on Selected Areas in Information Theory*, 1:157–166, 2020d.
- Y. Huang and A. C. Chung. Edge-variational graph convolutional networks for uncertainty-aware disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 562–572. Springer, 2020.
- S. L. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. Ranjit, A. Schwaighofer, F. Pérez-García, V. Salvatelli, S. Srivastav, A. Thieme, et al. MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv:2311.13668*, 2023.
- S. H. Hyun, M. S. Ahn, Y. W. Koh, and S. J. Lee. A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clinical Nuclear Medicine*, 44(12):956–960, 2019.
- IBM. CRISP-DM help overview — ibm.com. <https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview>, 2021. Accessed in Jun 2023.
- F. Imrie, S. Denner, L. S. Brunschwig, K. Maier-Hein, and M. van der Schaar. Automated ensemble multimodal machine learning for healthcare. *IEEE Journal of Biomedical and Health Informatics*, 2025.
- S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.
- I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Medical Imaging and Graphics*, 88:101843, 2021.
- J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, Jul. 2019.
- F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

- O. Israel, O. Pellet, L. Biassoni, D. De Palma, E. Estrada-Lobato, G. Gnanasegaran, T. Kuwert, C. La Fougère, G. Mariani, S. Massalha, et al. Two decades of SPECT/CT—the coming of age of a technology: An updated review of literature evidence. *European Journal of Nuclear Medicine and Molecular Imaging*, 46:1990–2012, 2019.
- M. Jin, M. T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv:1811.12276*, 2018.
- A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, and S. Horng. MIMIC-CXR-JPG-chest radiographs with structured labels. *PhysioNet*, 2019a.
- A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. MIMIC-CXR Database. <https://physionet.org/content/mimic-cxr/>, 2019b.
- A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark IV. MIMIC-IV (version 0.4). *PhysioNet*, 2020.
- A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV-Note: Deidentified free-text clinical notes, 2023a.
- A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and M. Roger. MIMIC-IV (version 3.0), 03 2024.
- A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023b.
- S. Jonas, A. O. Rossetti, M. Oddo, S. Jenni, P. Favaro, and F. Zubler. EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. *Human Brain Mapping*, 40(16):4606–4617, 2019.
- C. Jones, D. C. Castro, F. De Sousa Ribeiro, O. Oktay, M. McCradden, and B. Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146, 2024.
- G. Joshi, R. Walambe, and K. Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021.
- A. Kalapos and B. Gyires-Tóth. Self-supervised pretraining for 2D medical image segmentation. In *European Conference on Computer Vision*, pages 472–484. Springer, 2022.
- T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover

- regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.
- A. Kansal, E. Chen, T. Jin, P. Rajpurkar, and D. Kim. Multimodal clinical monitoring in the emergency department (MC-MED). *PhysioNet*, 1.0.0, 2025.
- O. Karaoğlu, H. Ş. Bilge, and İ. Uluer. Removal of speckle noises from ultrasound images using five different deep learning networks. *Engineering Science and Technology, an International Journal*, 29:101030, 2022.
- J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2018.
- S. P. Keehan, G. A. Cuckler, J. A. Poisal, A. M. Sisko, S. D. Smith, A. J. Madison, K. E. Rennie, J. A. Fiore, and J. C. Hardesty. National health expenditure projections, 2019–28: Expected rebound in prices drives rising spending growth. *Health Affairs*, 39(4):704–714, 2020.
- C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.
- F. Khader, J. N. Kather, G. Müller-Franzes, T. Wang, T. Han, S. Tayebi Arasteh, K. Hamesch, K. Bressen, C. Haarburger, J. Stegmaier, et al. Medical transformer for multimodal survival prediction in intensive care: Integration of imaging and non-imaging data. *Scientific Reports*, 13(1):10666, 2023.
- S. Khambampati, S. R. Dondapati, C. K. R. Bhuma, B. Madiraju, and R. K. Pathinarupothi. Frequency and time domain EEG analysis for prognostication of postanoxic comatose patients. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- M. O. Khan and Y. Fang. What is the best way to fine-tune self-supervised medical imaging models? In *Annual Conference on Medical Image Understanding and Analysis*, pages 267–281, 2024.
- S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.
- P. Kharazmi, S. Kalia, H. Lui, J. Wang, and T. Lee. A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. *Skin Research and Technology*, 24(2):256–264, 2018.
- D. Khazanova, V. C. Douglas, and E. Amorim. A matter of timing: EEG monitoring for neurological

- prognostication after cardiac arrest in the era of targeted temperature management. *Minerva Anestesiologica*, 87(6):704–713, 2021.
- D.-K. Kim, H.-C. Yoon, H.-S. Kim, W.-Y. Seo, and S.-H. Kim. Predicting neurological outcome after cardiac arrest using a pretrained model with electroencephalography augmentation. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- J. Kim, M. Chae, H.-J. Chang, Y.-A. Kim, and E. Park. Predicting cardiac arrest and respiratory failure using feasible artificial intelligence with simple trajectories of patient data. *Journal of clinical medicine*, 8(9):1336, 2019.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- D. G. Kirch and K. Petelle. Addressing the physician shortage: The peril of ignoring demography. *JAMA*, 317(19):1947–1948, 2017.
- H. Kittler, H. Pehamberger, K. Wolff, and M. Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165, 2002.
- J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807*, 2016.
- A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):1–14, 2022.
- W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13(10):818–829, 1985.
- R. P. Kocher. Reducing administrative waste in the US health care system. *JAMA*, 325(5):427–428, 2021.
- R. Krishnan, P. Rajpurkar, and E. J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, pages 1–7, 2022.
- C. Krittanawong, K. W. Johnson, R. S. Rosenson, Z. Wang, M. Aydar, U. Baber, J. K. Min, W. W. Tang, J. L. Halperin, and S. M. Narayan. Deep learning for cardiovascular medicine: A practical primer. *European Heart Journal*, 40(25):2058–2073, 2019.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.
- F. Krones and B. Walker. From theoretical models to practical deployment: A perspective and case study of opportunities and challenges in AI-driven cardiac auscultation research for low-income settings. *PLOS Digital Health*, 3(12):e0000437, 2024.

- F. Krones, B. Walker, G. Parsons, T. Lyons, and A. Mahdi. Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- F. Krones, B. Walker, T. Lyons, and A. Mahdi. Combining Hough transform and deep learning approaches to reconstruct ECG signals from printouts. *arXiv:2410.14185*, 2024.
- F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690, 2025.
- F. Krones, T. Papastylianou, G. Qian, G. Parsons, B. Shukla, B. W. Papież, and A. Mahdi. Advancing ICU decision-making through multimodal data fusion in predicting cardiovascular disorders. *Information Fusion*, 136:104573, 2026. doi: 10.1016/j.inffus.2026.104573.
- P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, et al. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *MedRxiv*, 2019.
- W. Lawson, G. Wagner, R. Startt-Selvester, and G. Ybarra. New method for digitization and computerized analysis of paper recordings of standard 12-lead electrocardiograms. In *Computers in cardiology 1995*, pages 41–44. IEEE, 1995.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- A. Lence, F. Extramiana, A. Fall, J.-E. Salem, J.-D. Zucker, and E. Prifti. Automatic digitization of paper electrocardiograms—a systematic review. *Journal of Electrocardiology*, 80:125–132, 2023.
- A. Leslie, A. Jones, and P. Goddard. The influence of clinical information on the reporting of CT by radiologists. *The British journal of radiology*, 73(874):1052–1055, 2000.
- K. Lewandowska, M. Weisbrot, A. Cieloszyk, W. Medrzycka-Dabrowska, S. Krupa, and D. Ozga. Impact of alarm fatigue on the work of nurses in an intensive care environment—a systematic review. *International journal of environmental research and public health*, 17(22):8409, 2020.
- C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, et al. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv:2205.12005*, 2022.
- H. Li and Y. Fan. Early prediction of Alzheimer’s disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 368–371. IEEE, 2019.

- M. Li, L. Xing, and A. J. Casson. Autoencoder artefact removal for brain signals and impact on classification performance. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- P. Li, Y. Hu, and Z.-P. Liu. Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods. *Biomedical Signal Processing and Control*, 66:102474, 2021.
- Q. Li, S. A. Spooner, M. Kaiser, N. Lingren, J. Robbins, T. Lingren, H. Tang, I. Solti, and Y. Ni. An end-to-end hybrid algorithm for automated medication discrepancy detection. *BMC Medical Informatics and Decision Making*, 15(1):1–12, 2015.
- S. Li, H. Shi, D. Sui, A. Hao, and H. Qin. A novel pathological images and genomic data fusion framework for breast cancer survival prediction. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1384–1387. IEEE, 2020a.
- Y. Li, Q. Qu, M. Wang, L. Yu, J. Wang, L. Shen, and K. He. Deep learning for digitizing highly noisy paper-based ECG records. *Computers in Biology and Medicine*, 127:104077, 2020b.
- J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. F. Williamson, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40(10):1095–1110, 2022.
- G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- Q. Liu and P. Hu. Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer. *Cancers*, 11(4):494, 2019.
- Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*, 2021.
- S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, and G. B. Kitchen. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, 38:4–9, 2021.
- J. M. Lopez Alcaraz and N. Strodthoff. MDS-ED: Multimodal decision support in the emergency department—a benchmark dataset for diagnoses and deterioration prediction in emergency medicine. *arXiv:2407.17856*, 2024.
- C. Lu, K. Chang, P. Singh, S. Pomerantz, S. Doyle, S. Kakarmath, C. Bridge, and J. Kalpathy-Cramer. Deploying clinical machine learning? consider the following... *arXiv:2109.06919*, 2021.

- A. M. Luks and E. R. Swenson. Pulse oximetry for monitoring patients with COVID-19 at home. Potential pitfalls and practical guidance. *Annals of the American Thoracic Society*, 17(9):1040–1046, 2020.
- S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
- M. N. Lyngbakken, P. L. Myhre, H. Røsjø, and T. Omland. Novel biomarkers of cardiovascular disease: Applications in clinical practice. *Critical reviews in clinical laboratory sciences*, 56(1):33–60, 2019.
- D. Ma, M. R. Hosseinzadeh Taher, J. Pang, N. U. Islam, F. Haghighi, M. B. Gotway, and J. Liang. Benchmarking and boosting transformers for medical image classification. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 12–22. Springer, 2022.
- K. Maharana, S. Mondal, and B. Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022.
- K. M. Martensen. *Radiographic image analysis-E-book*. Elsevier Health Sciences, 2013.
- H. Mayne, G. Parsons, and A. Mahdi. Unsupervised learning approaches for identifying ICU patient subgroups: Do results generalise? *arXiv:2403.02945*, 2024.
- N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.
- Y. Mei, G. Venkataramani, and T. Lan. Exploiting partial common information microstructure for multi-modal brain tumor segmentation. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 64–85. Springer, 2023.
- E. Merz and J. S. Abramowicz. 3D/4D ultrasound in prenatal diagnosis: Is it time for routine use? *Clinical Obstetrics and Gynecology*, 55(1):336–351, 2012.
- S. Mishra, G. Khatwani, R. Patil, D. Sapariya, V. Shah, D. Parmar, S. Dinesh, P. Daphal, and N. Mehendale. ECG paper record digitization and diagnosis using deep learning. *Journal of Medical and Biological Engineering*, 41(4):422–432, 2021.
- S. Mitra, M. Mitra, and B. Chaudhuri. Generation of digital time database from paper ECG records and Fourier transform-based analysis for disease identification. *Computers in biology and medicine*, 34(7):551–560, 2004.

- B. Mittelstadt, S. Wachter, and C. Russell. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *arXiv:2302.02404*, 2023.
- M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.
- J. Morley, C. C. Machado, C. Burr, J. Cows, I. Joshi, M. Taddeo, and L. Floridi. The ethics of AI in health care: A mapping review. *Social Science & Medicine*, 260:113172, 2020.
- C. Mosquera, L. Ferrer, D. H. Milone, D. R. Luna, and E. Ferrante. Understanding the impact of class imbalance on the performance of chest X-ray image classifiers. *arXiv:2112.12843*, 2021.
- N. A. Mullani, R. S. Herbst, R. G. O’Neil, K. L. Gould, B. J. Barron, and J. L. Abbruzzese. Tumor blood flow measured by PET dynamic imaging of first-pass ^{18}F -FDG uptake: A comparison with ^{15}O -labeled water-measured blood flow. *Journal of Nuclear Medicine*, 49(4):517–523, 2008.
- A. Mustafa and M. Rahimi Azghadi. Automated machine learning for healthcare and clinical notes analysis. *Computers*, 10(2):24, 2021.
- H. Namkoong, S. Yadlowsky, et al. Diagnosing model performance under distribution shift. *arXiv:2303.02011*, 2023.
- N. M. Nawi, W. H. Atomi, and M. Z. Rehman. The effect of data pre-processing on optimized training of artificial neural networks. *Procedia Technology*, 11:32–39, 2013.
- NBER. ICD-9-CM to and from ICD-10-CM and ICD-10-PCS crosswalk or general equivalence mappings — nber.org. <https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings>, 2024. Accessed in Oct 2024.
- M. Nevitt, D. Felson, and G. Lester. The osteoarthritis initiative. *Protocol for the Cohort Study*, 1, 2006.
- NHS. Diagnostic imaging dataset statistical release. *London: Department of Health*, 421, 2016.
- A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti. The importance of respiratory rate monitoring: From healthcare to sport and exercise. *Sensors*, 20(21):6396, 2020.
- D. Nie, J. Lu, H. Zhang, E. Adeli, J. Wang, Z. Yu, L. Liu, Q. Wang, J. Wu, and D. Shen. Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific Reports*, 9(1):1–14, 2019.
- Y. Ning, S. Teixayavong, Y. Shang, J. Savulescu, V. Nagaraj, D. Miao, M. Mertens, D. S. W. Ting, J. C. L. Ong, M. Liu, J. Cao, M. Dunn, R. Vaughan, M. E. H. Ong, J. J.-Y. Sung, E. J. Topol,

- and N. Liu. Generative artificial intelligence in healthcare: Ethical considerations and assessment checklist. *arXiv:2311.02107*, 2023.
- M. Nishimori, K. Kiuchi, K. Nishimura, K. Kusano, A. Yoshida, K. Adachi, Y. Hirayama, Y. Miyazaki, R. Fujiwara, P. Sommer, et al. Accessory pathway analysis using a multimodal deep learning model. *Scientific Reports*, 11(1):1–8, 2021.
- H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv:2311.16452*, 2023.
- Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- S. Oh, N. Kim, and J. Ryu. Analyzing to discover origins of CNNs and ViT architectures in medical images. *Scientific Reports*, 14(1):8755, 2024.
- OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.
- OpenAI. GPT-4o system card. <https://cdn.openai.com/gpt-4o-system-card.pdf>, 2024.
- A. Palepu and A. L. Beam. TIER: Text-image entropy regularization for CLIP-style models. *arXiv:2212.06710*, 2022.
- J. Pang, F. Haghighi, D. Ma, N. U. Islam, M. R. Hosseinzadeh Taher, M. B. Gotway, and J. Liang. POPAR: Patch order prediction and appearance recovery for self-supervised medical image analysis. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 77–87. Springer, 2022.
- S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer’s disease. *Medical Image Analysis*, 48:117–130, 2018.
- S. Patidar and N. K. Sawant. A tensor decomposition-based feature extraction method to predict neurological recovery from coma after cardiac arrest using EEG signals. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

- K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007.
- X. Pei, K. Zuo, Y. Li, and Z. Pang. A review of the application of multi-modal deep learning in medicine: Bibliometrics and future directions. *International Journal of Computational Intelligence Systems*, 16(1):44, 2023.
- S. R. Pfohl, Y. Xu, A. Foryciarz, N. Ignatiadis, J. Genkins, and N. H. Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. *arXiv:2202.01906*, 2022.
- Philips Healthcare. IntelliVue Guardian Early Warning Score (EWS) Solution — philips.de. <https://www.philips.de/healthcare/product/HCN0CTN60/intellivue-guardian-early-warning-score-ews-guardian-lsung>, 2026. Accessed in Feb 2026.
- F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, and G. Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.
- E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.
- M. R. Pinsky, A. Bedoya, A. Bihorac, L. Celi, M. Churpek, N. J. Economou-Zavlanos, P. Elbers, S. Saria, V. Liu, P. G. Lyons, et al. Use of artificial intelligence in critical care: Opportunities and obstacles. *Critical Care*, 28(1):113, 2024.
- D. Placido, H.-C. Thorsen-Meyer, B. S. Kaas-Hansen, R. Reguant, and S. Brunak. Development of a dynamic prediction model for unplanned ICU admission and mortality in hospitalized patients. *PLOS digital health*, 2(6):e0000116, 2023.
- R. A. Poldrack, K. J. Gorgolewski, and G. Varoquaux. Computational and informatic advances for reproducible data analysis in neuroimaging. *Annual Review of Biomedical Data Science*, 2:119–138, 2019.
- S. Pölsterl, T. N. Wolf, and C. Wachinger. Combining 3D image and tabular data via the dynamic affine feature map transform. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 688–698. Springer, 2021.
- S. J. Prince. *Understanding deep learning*. MIT Press, 2023. URL <http://udlbook.com>.
- S. Purwar, R. K. Tripathi, R. Ranjan, and R. Saxena. Detection of microcytic hypochromia using cbc

- and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications*, 79(7):4573–4595, 2020.
- A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180, 2020.
- S. Qiu, G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama. Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:737–749, 2018.
- S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino. Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion*, 80:241–265, 2022.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. AI in health and medicine. *Nature Medicine*, pages 1–8, 2022.
- V. Raleigh, D. Jefferies, and D. Wellings. Cardiovascular disease in England. *The King’s Fund*, pages 2022–11, 2022.
- D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.
- L. Ravichandran, C. Harless, A. J. Shah, C. A. Wick, J. H. McClellan, and S. Tridandapani. Novel tool for complete digitization of paper electrocardiography data. *IEEE journal of translational engineering in health and medicine*, 1:1800107–1800107, 2013.
- I. Reda, A. Khalil, M. Elmogy, A. Abou El-Fetouh, A. Shalaby, M. Abou El-Ghar, A. Elmaghraby, M. Ghazal, and A. El-Baz. Deep learning role in early diagnosis of prostate cancer. *Technology in Cancer Research & Treatment*, 17:1533034618775530, 2018.
- J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädtsch,

- C. H. Sudre, L. Acion, M. Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, pages 1–13, 2024.
- P. Renc, Y. Jia, A. E. Samir, J. Was, Q. Li, D. W. Bates, and A. Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7:256, 2024.
- M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. v. Tengg-Kobligk, R. M. Summers, and R. Wiest. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.
- M. A. Reyna, Y. Kiarashi, A. Elola, J. Oliveira, F. Renna, A. Gu, E. A. P. Alday, N. Sadr, A. Sharma, J. Kpodonu, S. Mattos, M. T. Coimbra, R. Sameni, A. B. Rad, and G. D. Clifford. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. *medRxiv*, 2022.
- M. A. Reyna, E. Amorim, R. Sameni, J. Weigle, A. Elola, A. Bahrami, S. Seyedi, H. Kwon, W.-L. Zheng, M. Ghassemi, M. J. A. M. van Putten, J. Hofmeijer, N. Gaspard, A. Sivaraju, S. Herman, J. W. Lee, B. M. Westover, and G. D. Clifford. Predicting neurological recovery from coma after cardiac arrest: The George B. Moody PhysioNet challenge 2023. *Computing in Cardiology (CinC)*, 50:1–4, 2023.
- M. A. Reyna, Deepanshi, J. Weigle, Z. Koscova, K. Campbell, K. Kodthalu Shivashankara, S. Saghafi, S. Nikookar, M. Motie-Shirazi, Y. Kiarashi, S. Seyedi, G. D. Clifford, and R. Sameni. ECG-Image-Database: A dataset of ECG images with real-world imaging and scanning artifacts; a foundation for computerized ECG image digitization and analysis. *arXiv:2409.16612*, 2024a.
- M. A. Reyna, Deepanshi, J. Weigle, Z. Koscova, A. Elola, S. Seyedi, K. Campbell, G. D. Clifford, and R. Sameni. Digitization and classification of ECG images: The George B. Moody PhysioNet challenge. *Computing in Cardiology (CinC)*, 51:1–4, 2024b.
- T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor. ImageNet-21k pretraining for the masses. *arXiv:2104.10972*, 2021.
- N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. The future of digital health with federated learning. *npj Digital Medicine*, 119(3):1–7, 2020.
- I. Rodin, I. Fedulova, A. Shelmanov, and D. V. Dylov. Multitask and multimodal neural network model for interpretable analysis of X-Ray images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1601–1604. IEEE, 2019.
- E. Rubinstein, M. Salhov, M. Nidam-Leshem, V. White, S. Golan, J. Baniel, H. Bernstine, D. Groshar,

- and A. Averbuch. Unsupervised tumor detection in Dynamic PET/CT imaging of the prostate. *Medical Image Analysis*, 55:27–40, 2019.
- B. J. Ruijter, M. C. Tjepkema-Cloostermans, S. C. Tromp, W. M. van den Bergh, N. A. Foudraïne, F. H. Kornips, G. Drost, E. Scholten, F. H. Bosch, A. Beishuizen, et al. Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study. *Annals of neurology*, 86(2): 203–214, 2019.
- S. G. Ryman and K. L. Poston. MRI biomarkers of motor and non-motor symptoms in Parkinson’s disease. *Parkinsonism & Related Disorders*, 73:85–93, 2020.
- J. Rystrom, Z. Fu, and C. Russell. OxEnsemble: Efficient fair ensembles for low-data classification. In *Submitted to Medical Imaging with Deep Learning*, 2025. URL <https://openreview.net/forum?id=DURUqZgwK8>. under review.
- B. Sahiner, W. Chen, R. K. Samala, and N. Petrick. Data drift in medical machine learning: Implications and potential remedies. *The British Journal of Radiology*, page 20220878, 2023.
- M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Computers in Biology and Medicine*, 129:104150, 2021.
- M. Salvi, H. W. Loh, S. Seoni, P. D. Barua, S. García, F. Molinari, and U. R. Acharya. Multi-modality approaches for medical support systems: A systematic review of the last decade. *Information Fusion*, page 102134, 2023.
- Z. A. Samak, P. Clatworthy, and M. Mirmehdi. Prediction of thrombectomy functional outcomes using multimodal data. In *Annual Conference on Medical Image Understanding and Analysis*, pages 267–279. Springer, 2020.
- A. Saporta, A. M. Puli, M. Goldstein, and R. Ranganath. Symile-MIMIC: A multimodal clinical dataset of chest X-rays, electrocardiograms, and blood labs from MIMIC-IV. PhysioNet, version 1.0.0, 2025.
- J. Schoepf, P. Zwerner, G. Savino, C. Herzog, J. M. Kerl, and P. Costello. Coronary CT angiography. *Radiology*, 244(1):48–63, 2007.
- S. Schulz, A. Woerl, F. Jungmann, C. Glasner, P. Stenzel, S. Strobl, A. Fernandez, D. Wagner, A. Haferkamp, P. Mildenberger, et al. Multimodal deep learning for prognosis prediction in renal cancer. *Frontiers in Oncology*, 11, 2021.
- J. Sebastian and G. G. King. Fusion of multimodality medical images—A review. *2021 Smart Technologies, Communication and Robotics (STCR)*, pages 1–6, 2021.

- N. A. Segal, M. C. Nevitt, K. D. Gross, J. Hietpas, N. A. Glass, C. E. Lewis, and J. C. Torner. The Multicenter Osteoarthritis Study (MOST): Opportunities for rehabilitation research. *PM & R: the Journal of Injury, Function, and Rehabilitation*, 5(8), 2013.
- A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.
- 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.
- L. Seyyed-Kalantari, H. Zhang, M. McDermott, I. Y. Chen, and M. Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, pages 1–7, 2021.
- T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, page 102040, 2023.
- S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, 2021.
- E. Sheetrit, M. Brief, and O. Elisha. Predicting unplanned readmissions in the intensive care unit: A multimodality evaluation. *Scientific Reports*, 13(1):15426, 2023.
- S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani, et al. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, 7(2): e12239, 2019.
- D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.
- K. K. Shivashankara, Deepanshi, A. M. Shervedani, M. A. Reyna, G. D. Clifford, and R. Sameni. ECG-Image-Kit: A synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. *Physiological Measurement*, 45:055019, 2024.
- S. Shurrab and R. Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

- R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.
- L. A. V. Silva and K. Rohr. Pan-cancer prognosis prediction using multimodal deep learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 568–571. IEEE, 2020.
- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.
- V. Sivaraman, L. A. Bukowski, J. Levin, J. M. Kahn, and A. Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.
- M. E. B. Smith, J. C. Chiovaro, M. O’Neil, D. Kansagara, A. R. Quiñones, M. Freeman, M. L. Motu’apuaka, and C. G. Slatore. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Annals of the American Thoracic Society*, 11(9):1454–1465, 2014.
- I. Sobel. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, 02 2014.
- L. R. Soenksen, Y. Ma, C. Zeng, L. D. Boussieux, K. V. Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *arXiv:2202.12998*, 2022.
- H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar. Moco pretraining improves representation and transferability of chest X-Ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021.
- I. Spasic, G. Nenadic, et al. Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics*, 8(3):e17984, 2020.
- S. E. Spasov, L. Passamonti, A. Duggento, P. Liò, and N. Toschi. A multi-modal convolutional neural network framework for the prediction of Alzheimer’s disease. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1271–1274. IEEE, 2018.
- A. Stadler, W. Schima, A. Ba-Ssalamah, J. Kettenbach, and E. Eisenhuber. Artifacts in body MR imaging: Their appearance and how to eliminate them. *European Radiology*, 17:1242–1255, 2007.
- S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren. Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics*, 2022.
- E. Stenhede, B.-J. Singstad, and A. Ranjbar. ECG-DUaL: Pose-invariant digitization of printed

- electrocardiograms via U-Net and local edge detection. In *2024 Computing in Cardiology (CinC)*, 2024.
- N. Strodthoff, J. M. Lopez Alcaraz, and W. Haverkamp. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: An explorative study in emergency care. *European Heart Journal-Digital Health*, 5(4):454–460, 2024.
- H.-I. Suk, S.-W. Lee, D. Shen, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.
- R. Summers. NIH chest X-Ray dataset of 14 common thorax disease categories, 2019.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- A. Taleb, C. Lippert, T. Klein, and M. Nabi. Multimodal self-supervised learning for medical image analysis. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings*, pages 661–673. Springer, 2021.
- L. K. Tan. Image file formats. *Biomed Imaging Interv J*, 2(1):e6, 2006.
- M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- D. Thiele, R. Rodseth, R. Friedland, F. Berger, C. Mathew, C. Maslo, V. Moll, C. Leithner, C. Storm, A. Krannich, et al. Machine learning models for the early real-time prediction of deterioration in intensive care units—a novel approach to the early identification of high-risk patients. *Journal of Clinical Medicine*, 14(2):350, 2025.
- K.-H. Thung, P.-T. Yap, and D. Shen. Multi-stage diagnosis of Alzheimer’s disease with incomplete multimodal data via multi-task deep learning. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 160–168. Springer, 2017.
- T. Tirupal, B. C. Mohan, and S. S. Kumar. Multimodal medical image fusion techniques—A review. *Current Signal Transduction Therapy*, 16(2):142–163, 2021.
- G. H. Tison, J. Zhang, F. N. Delling, and R. C. Deo. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. *Circulation: Cardiovascular Quality and Outcomes*, 12(9):e005289, 2019.
- E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of

- pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022.
- A. Tiulpin, S. Klein, S. M. Bierma-Zeinsträ, J. Thevenot, E. Rahtu, J. van Meurs, E. H. Oei, and S. Saarakkala. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific Reports*, 9(1):1–11, 2019.
- M. C. Tjepkema-Cloostermans, C. da Silva Lourenço, B. J. Ruijter, S. C. Tromp, G. Drost, F. H. Kornips, A. Beishuizen, F. H. Bosch, J. Hofmeijer, and M. J. van Putten. Outcome prediction in postanoxic coma with deep learning. *Critical Care Medicine*, 47(10):1424–1432, 2019.
- E. Tjoa and C. Guan. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- E. J. Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.
- E. J. Topol. As artificial intelligence goes multimodal, medical applications multiply, 2023.
- K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Human language technologies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.
- D. Tran, J. Liu, M. W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, et al. PLEX: Towards reliability using pretrained large model extensions. *arXiv:2207.07411*, 2022.
- P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.
- T. Tu, S. Azizi, D. Driess, M. Schaeckermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical AI. *arXiv:2307.14334*, 2023.
- F. Uslenghi, R. Sassi, and M. W. Rivolta. Recovery from coma after cardiac arrest: Which time-window counts the most for deep learning predictions? In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- M. Van Der Schaar and A. Rashbass. The case for reality-centric AI — vanderschaar-lab.com. <https://www.vanderschaar-lab.com/the-case-for-reality-centric-ai/>, 2023. Accessed in Jan 2024.
- R. S. Vanguri, J. Luo, A. T. Aukerman, J. V. Egger, C. J. Fong, N. Horvat, A. Pagano, J. d. A. B. Araujo-Filho, L. Geneslaw, H. Rizvi, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. *Nature Cancer*, 3(10):1151–1164, 2022.

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- R. Vavekanand, T. Kumar, S. Kumar, G. Kumar, and A. A. Laghari. Multimodal machine learning approaches in predictive healthcare analytics: a comprehensive survey. *Archives of Computational Methods in Engineering*, pages 1–25, 2026.
- J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of Alzheimer’s disease stage. *Scientific Reports*, 11(1):1–13, 2021.
- M. A. Verlyck, J. R. Dillon, S. A. Creamer, and D. Zhao. WAVIE: A modular and open-source Python implementation for fully automated digitisation of paper electrocardiograms. In *2024 Computing in Cardiology (CinC)*, 2024.
- S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare ’18*, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.
- M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: A meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, 159(3):669–676, 2008.
- J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix). *Intensive care medicine*, 22(7):707–710, 1996.
- K. Vishwanath, A. Alyakin, M. Ghosh, A. Hage, S. N. Neifert, C. Orillac, N. J. Mandelberg, H. A. Khan, J. V. Lee, J. J. Yao, W. R. Small, A. Varma, D. B. Hewitt, Y. Aphinyanaphongs, D. A. Alber, and E. K. Oermann. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nature Medicine*, 2026. doi: 10.1038/s41591-026-04431-5.
- Viz.ai. Viz.ai – AI-Powered Care Coordination Platform — viz.ai. <https://www.viz.ai/>, 2026. Accessed in Feb 2026.
- P. Voigt and A. Von dem Bussche. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.
- S. Wachter, B. Mittelstadt, and C. Russell. Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

- S. Wachter, B. Mittelstadt, and C. Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41:105567, 2021.
- P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter. PTB-XL: A large publicly available ECG dataset. *Scientific Data*, 2020.
- G. S. Waits and E. Z. Soliman. Digitizing paper electrocardiograms: Status and challenges. *Journal of Electrocardiology*, 50(1):123–130, 2017.
- B. Walker, F. Krones, I. Kiskin, G. Parsons, T. Lyons, and A. Mahdi. Dual Bayesian ResNet: A deep learning approach to heart murmur detection. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4, 2022.
- J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.
- G. Wang, X. Liu, Z. Ying, G. Yang, Z. Chen, Z. Liu, M. Zhang, H. Yan, Y. Lu, Y. Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: A proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.
- J. Wang, H. Zhu, S.-H. Wang, and Y.-D. Zhang. A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26(1):351–380, 2021a.
- J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan. OmniVL: One foundation model for image-language and video-language tasks. *arXiv:2209.07526*, 2022a.
- J. T. Wang and D. P. Mital. A microcomputer-based prototype for ECG paper record conversion. *Journal of network and computer applications*, 19(3):295–307, 1996.
- S. Wang, Y. Yin, D. Wang, Y. Wang, and Y. Jin. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Transactions on Cybernetics*, 52(12):12623–12637, 2021b.
- X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

- Z. Wang, Y. Bai, Y. Zhou, and C. Xie. Can CNNs be more robust than transformers? *arXiv:2206.03452*, 2022b.
- Z. Wang, Y. Wang, H. Zhang, W. Wang, J. Qi, J. Chen, N. Sastry, J. Johnson, and S. De. ICDXML: Enhancing ICD coding with probabilistic label trees and dynamic semantic representations. *Scientific Reports*, 14(1):18319, 2024.
- R. R. Watkins and T. L. Lemonovich. Role of inflammatory markers in the diagnosis and management of infective endocarditis. *Infectious Diseases in Clinical Practice*, 18(2):87–90, 2010.
- D. Watson. *Explaining black box algorithms: Epistemological challenges and machine learning solutions*. PhD thesis, University of Oxford, 2021.
- M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack Jr, W. Jagust, J. C. Morris, et al. The Alzheimer’s Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. *Alzheimer’s & Dementia*, 13(5):561–571, 2017.
- Y. Weng, Y. Zhang, W. Wang, and T. Dening. Semi-supervised information fusion for medical image analysis: Recent progress and future perspectives. *Information Fusion*, page 102263, 2024.
- L. E. Widman and L. S. Hines. Digitization of electrocardiograms by desktop optical scanner. *Journal of electrocardiology*, 24(4):325–338, 1991.
- K. Widner, S. Virmani, J. Krause, J. Nayar, R. Tiwari, E. R. Pedersen, D. Jeji, N. Hammel, Y. Matias, G. S. Corrado, et al. Lessons learned from translating AI from development to deployment in healthcare. *Nature Medicine*, pages 1–3, 2023.
- R. S. Wiener, L. M. Schwartz, and S. Woloshin. When a test is too good: How CT pulmonary angiograms find pulmonary emboli that do not need to be found. *BMJ*, 347, 2013.
- M. A. Wójcik. Foundation models in healthcare: Opportunities, biases and regulatory prospects in Europe. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 32–46. Springer, 2022.
- J. Woo. A short history of the development of ultrasound in obstetrics and gynecology. *History of Ultrasound in Obstetrics and Gynecology*, 3:1–25, 2002.
- World Health Organisation. Cardiovascular diseases (CVDs) — WHO.int. [https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)), 2021. Accessed in May 2023.
- World Health Organisation. *World health statistics 2023: Monitoring health for the SDGs, Sustainable Development Goals*. Geneva: World Health Organisation, 2023.

- H. Wu, K. H. K. Patel, X. Li, B. Zhang, C. Galazis, N. Bajaj, A. Sau, X. Shi, L. Sun, Y. Tao, et al. A fully-automated paper ECG digitisation algorithm using deep learning. *Scientific Reports*, 12(1): 20963, 2022.
- K. Wu, E. Wu, B. Theodorou, W. Liang, C. Mack, L. Glass, J. Sun, and J. Zou. Characterizing the clinical adoption of medical AI devices through US insurance claims. *NEJM AI*, 2023.
- W. Wu and Y. Tan. MelicientNet: Harnessing Mel-spectrograms and EfficientNet architectures for predicting neurological recovery post-cardiac arrest. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- P. Xia, D. Zhao, Y. Yao, Z. Bai, Y. Shao, S. Lu, F. Geng, Y. Zhu, P. Wang, and L. Du. Predicting neurological outcomes for cardiac arrest patients from longitudinal EEG based on short-time Fourier transform and 3-D deep residual network. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, page 101985, 2021.
- H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv:2302.00402*, 2023a.
- M. Xu, L. Ouyang, Y. Gao, Y. Chen, T. Yu, Q. Li, K. Sun, F. S. Bao, L. Safarnejad, J. Wen, et al. Accurately differentiating COVID-19, other viral infection, and healthy individuals using multimodal features via late fusion learning. *medRxiv*, 2020.
- X. Xu, C. Wu, S. Rosenman, V. Lal, W. Che, and N. Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647, 2023b.
- Z. Xu, D. R. So, and A. M. Dai. Mufasa: Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10532–10540, 2021.
- A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66, 2019.
- R. Yan, F. Zhang, X. Rao, Z. Lv, J. Li, L. Zhang, S. Liang, Y. Li, F. Ren, C. Zheng, et al. Richer fusion network for breast cancer classification based on multimodal data. *BMC Medical Informatics and Decision Making*, 21(1):1–15, 2021.

- H. Yang and R. Tetzlaff. Model ensembling for predicting neurological recovery after cardiac arrest: Top-down or bottom-up? In *2023 Computing in Cardiology (CinC)*, pages 1–4, 2023.
- Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, pages 1–11, 2024.
- J. Yap, W. Yolland, and P. Tschandl. Multimodal skin lesion classification using deep learning. *Experimental Dermatology*, 27(11):1261–1267, 2018.
- Yaseen, G.-Y. Son, and S. Kwon. Classification of heart sound signal using multiple features. *Applied Sciences*, 8(12), 2018.
- N. Yildirim, H. Richardson, M. T. Wetscherek, J. Bajwa, J. Jacob, M. A. Pinnock, S. Harris, D. C. de Castro, S. Bannur, S. L. Hyland, et al. Multimodal healthcare AI: Identifying and designing clinically relevant vision-language applications for radiology. *arXiv:2402.14252*, 2024.
- Y. Yoo, L. Y. Tang, D. K. Li, L. Metz, S. Kolind, A. L. Traboulsee, and R. C. Tam. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(3):250–259, 2019.
- H.-C. Yoon, D.-K. Kim, H.-S. Kim, W.-Y. Seo, C.-H. Heo, and S.-H. Kim. Segmentation-based extraction of key components from ECG images: A framework for precise classification and digitization. In *2024 Computing in Cardiology (CinC)*, 2024.
- J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk. Domain generalization for medical image analysis: A review. *arXiv:2310.08598*, 2023.
- D. Yu, G. Sheikholeslami, and A. Zhang. Findout: Finding outliers in very large datasets. *Knowledge and information Systems*, 4:387–412, 2002.
- X. Yu, Y. Huang, J. Wu, J. Wang, and W. Cai. From paper to digital: ECG processing with U-Net digitization and ResNet classification. In *2024 Computing in Cardiology (CinC)*, 2024.
- M. Zabihi, A. C. Zar, P. Grover, and E. S. Rosenthal. Hyperensemble learning from multimodal biosignals to robustly predict functional outcome after cardiac arrest. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023.
- M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.
- H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021.

- S. Zhang, S. M. H. Bamakan, Q. Qu, and S. Li. Learning for personalized medicine: A comprehensive review from a deep learning perspective. *IEEE Reviews in Biomedical Engineering*, 12:194–208, 2018.
- T. Zhao, E. Dai, K. Shu, and S. Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, page 1433–1442, New York, NY, USA, 2022. Association for Computing Machinery.
- W.-L. Zheng, E. Amorim, J. Jing, W. Ge, S. Hong, O. Wu, M. Ghassemi, J. W. Lee, A. Sivaraju, T. Pang, S. T. Herman, N. Gaspard, B. J. Ruijter, J. Sun, M. C. Tjepkema-Cloostermans, J. Hofmeijer, M. J. van Putten, and M. B. Westover. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. *Resuscitation*, 169:86–94, 2021.
- J. Zhou, X. Zhang, Z. Zhu, X. Lan, L. Fu, H. Wang, and H. Wen. Cohesive multi-modality feature learning and fusion for COVID-19 patient severity prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2535–2549, 2021a.
- S. Zhou, S. Geng, J. Li, D. Zhang, Z. Xie, C. Cheng, and S. Hong. Less is more: Reducing overfitting in deep learning for EEG classification. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4, 2023a.
- T. Zhou, K.-H. Thung, X. Zhu, and D. Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human Brain Mapping*, 40(3): 1001–1016, 2019.
- T. Zhou, Q. Cheng, H. Lu, Q. Li, X. Zhang, and S. Qiu. Deep learning methods for medical image fusion: A review. *Computers in Biology and Medicine*, page 106959, 2023b.
- Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, et al. RadFusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR. *arXiv:2111.11665*, 2021b.
- Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, pages 1–8, 2023c.
- H. Zhu and S. Wang. Learning fair models without sensitive attributes: A generative approach. *arXiv:2203.16413*, 2022.
- F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

- D. Zietlow, M. Lohaus, G. Balakrishnan, M. Kleindessner, F. Locatello, B. Schölkopf, and C. Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.
- J. E. Zimmerman, A. A. Kramer, D. S. McNair, F. M. Malila, and V. L. Shaffer. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Critical care medicine*, 34(10):2517–2529, 2006.