Interpretable Machine Learning for Predictive Analytics with High-Fidelity Imbalanced Clinical Data

Suraj N. Ramchand
[redacted]

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

June 23, 2025

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .....[redacted]..... (candidate)

Date .....23/06/2025.....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ...[redacted]..... (candidate)

Date .....23/06/2025.....

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..[redacted]..... (candidate)

Date .....23/06/2025.....

This work is dedicated to my wife, family, and friends, especially to Fares Abdullah, who was a tremendous source of support and encouragement throughout this journey. I extend my deepest gratitude to all who have been pillars of love and strength, filling my days with joy and purpose. I also honour those who have left this world, leaving a legacy of memories and lessons that continue to guide me. Each of you has shaped my story in ways words can scarcely capture. Your impact resonates on every page of this work.

Abstract

Healthcare professionals frequently manage complex, under-represented clinical events that challenge established diagnostic and decision-making pathways. These include infrequently diagnosed conditions, early-stage deterioration, and diseases with heterogeneous manifestations, where low prevalence, variable presentation, and limited structured data introduce significant uncertainty. Clinicians apply deep expertise to interpret these cases, often under constraints of time, information, and documentation. This thesis examines how high-fidelity clinical data can be analysed using machine learning (ML) techniques to support clinical judgement. It explores how computational models might surface latent patterns, identify early risk signals, and make complex data more interpretable in settings defined by imbalance, variability, and diagnostic ambiguity.

Across a series of empirical studies, the thesis addresses technical challenges common to rare or complex event prediction: class imbalance, missingness, temporal variation, and heterogeneity in clinical presentation. The first case study focuses on sepsis, a common but acutely time-sensitive condition. While sepsis is well recognised, the rapid progression of the disease limits the availability of temporally labelled pre-onset data. Using intensive care unit (ICU) records, the thesis develops preprocessing pipelines and tests attention-based temporal models to support early warning in a moderately imbalanced context, where timely prediction is critical and observational data are sparse.

With the advent of COVID-19, a fast-spreading and clinically disruptive condition, understanding the factors that lead to hospitalisation became a pressing research need. While early studies primarily focused on symptom onset and complications, less attention was paid to the early signals embedded in longitudinal health records. This study draws on primary care data that capture patients' health trajectories over time, enabling analysis of events leading up to hospitalisation. Within this broader population dataset, hospital admission occurs in approximately 10% of cases, resulting in a naturally imbalanced outcome. Class-sensitive objectives and temporally structured features are applied to surface early risk markers that may inform

triage decisions and support timely intervention.

Building on the hospitalisation study, the thesis further explores how injecting structured clinical knowledge into models can improve the representation of infrequent manifestations within the same imbalanced primary care setting. Medical ontologies are integrated into language models to enhance the embedding of rare clinical terms and improve classification performance. This approach is applied to two additional prediction tasks using the same dataset. COVID-19-related mortality (1%) serves as a highly imbalanced clinical outcome for evaluating model sensitivity to low-signal, high-risk events, whilst stroke (30%) is used as a more common benchmark for assessing generalisability across conditions of varying prevalence and complexity. Together, these tasks extend the modelling framework to uncover how structured knowledge improves the encoding of infrequent clinical features and contributes to more robust and generalisable prediction across diverse diagnostic contexts.

Grounded in the clinical utility of differentiating hypertrophic cardiomyopathy (HCM) from Fabry disease, a rare metabolic disorder frequently misdiagnosed due to overlapping cardiac features, this final case study extends the thesis's exploration of imbalance to include disease heterogeneity and population-level rarity. A novel multimodal dataset was collected from hospital cardiac records, including echocardiography, ECG, and Holter monitoring data from genetically confirmed Fabry patients and matched HCM controls. Traditional ML classifiers trained on this dataset showed promising discriminatory ability using routinely acquired clinical measurements. These findings suggest that standard diagnostic tools may help raise earlier consideration of rare conditions in everyday practice when modelled with care and clinical insight.

Together, these studies propose a generalisable approach to modelling rare and underrepresented clinical events using ML. The thesis focuses on structuring and improving the quality and utility of clinical data, surfacing patterns of clinical relevance, and supporting decision-making in diagnostically uncertain environments. The final clinical interface is designed as a practical tool to assist interpretation and integrate predictions into clinical workflows and as a space to understand better how models behave in context. These contributions are made with deep appreciation for the complexity of clinical decision-making and the expertise of those who carry it out. The work aims to support that expertise by offering tools and insights that are transparent, interpretable, and aligned with real-world care.

Acknowledgements

Reflecting on my PhD journey at Swansea, I am grateful for the incredible support and guidance I have received from many individuals. Pursuing a doctoral degree is never easy, and the challenges were only compounded by the global pandemic that disrupted our lives in unprecedented ways. Yet, through it all, I have been fortunate to have mentors, colleagues, friends, and family who have stood by me, lifted me up, and helped me reach this milestone.

First and foremost, my heartfelt gratitude goes to my wife and family. Your unwavering support and belief in me have been the bedrock of my perseverance and success. My dear partner, thank you for the endless patience, love, and encouragement during the most challenging moments. To my family, your sacrifices, understanding, and boundless love have moulded me into the person I am today. This accomplishment is as much yours as it is mine.

I must also express my profound appreciation to my primary supervisor, Xianghua Xie. Your guidance, patience, and wisdom have been instrumental in shaping me into the researcher I am today. Thank you for believing in me, even when I doubted myself, and for pushing me to strive for excellence in my work. I am equally grateful to my second supervisor, Duncan Cole, whose insightful feedback and encouragement have been invaluable. Your guidance has helped me navigate my research complexities and greatly enriched my work. I am also thankful to the clinicians who supported this work, especially Avraj Virdi and Professor Yousef, for their generous input and thoughtful clinical perspectives.

To my colleagues in the CSVision research group, thank you for creating such a supportive and collaborative environment. Our academic and personal discussions have been a source of inspiration and have made this journey so much more enjoyable.

I thank my friends, who have been my companions on this journey, offering their support, laughter, and invaluable moments of respite. Your camaraderie and unwavering belief in my abilities have strengthened and motivated me. I am truly blessed to have you in my life.

Last but not least, I am grateful to Amicus Therapeutics UK Operations and the Engineer-

ing and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Enhancing Human Interactions and Collaborations with Data and Intelligence-Driven Systems, grant number EP/S021892/1, for their support. Your faith in my research has enabled me to pursue this path with confidence and dedication.

To everyone who has been a part of this journey, whether mentioned here or not, please know that your contributions, big or small, have profoundly impacted my life and work. I am forever grateful for your support, and I share this achievement with all of you. Thank you from the bottom of my heart.

Contents

Abbreviations

RNN	Recurrent Neural Network
LSTM	Long-Short Term Memory
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
VAE	Variational Auto-Encoder
COVID-19	Coronavirus Disease 2019
NGS	Next-generation sequencing
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
UK	United Kingdom
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic curve
GP	General Practice
1D-CNN	1D Convolutional Neural Network
SCCS	Self-Controlled Case Series
NHS	National Health Service
L2D	Learn to Diagnose
ESM	Encounter Sequence Modelling
Bi-LSTM	Bidirectional Long-Short Term Memory
Retain	REverse Time AttentIoN model

RetainEX	REverse Time AttentIoN EX model
RetainEXT	REverse Time AttentIoN EXTension model
GI	Gastrointestinal
PDT	Physician Diagnostic Thinking
RF	Random Forest
GB	Gradient Boosting
XGB	eXtreme Gradient Boosting
TNN	Temporal Neural Networks
FBC	Full Blood Count
GNN	Graph Neural Networks
GCN	Graph Convolutional Networks
DAG	Directed Acyclic Graph
GAT	Graph Attention Networks
GATv2Conv	Graph Attention v2 Convolution
NLP	Natural Language Processing
LLM	Large Language Model
LLMs	Large Language Models
BEHRT	Bidirectional Encoder Electronic Health Records Representations from Transformers
SAIL	Secure Anonymised Information Linkage
CAM	Class Activation Maps
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
VBHC	Value-Based Healthcare
ICU	Intensive Care Unit
LOS	Lenght Of Stay
NCL	Neighbourhood Cleaning Rule

MIMIC-IV	Medical Information Mart for Intensive Care-IV
GCS	Glasgow Comma Scale
PDPs	Partial Dependency Plots
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	SHapley Additive exPlanations
IDx-DR	IDx-DR diabetic retinopathy diagnostic system
EWS	Early Warning Score
FiO_2	Fraction of Inspired Oxygen
PaO_2	Partial Pressure of Oxygen
SOFA	Sequential Organ Failure Assessment
SIRS	Systemic Inflammatory Response Syndrome
FD	Fabry Disease
HCM	Hypertrophic cardiomyopathy
ECG	Electrocardiogram
GLA	α-galactosidase A gene
PPV	Positive Predictive Value
SHAP	SHapley Additive exPlanations
PDP	Partial Dependency Plots
SGD	Stochastic Gradient Descent
LVPWd	Left Ventricular Pulse Wave Duration
KNN	K-Nearest Neighbors
LVH	Left Ventricular Hypertrophy
ET	Endothelin
MLP	Multi-Layer Perceptron
LVIDs	Left Ventricular Internal Dimension at Systole
IVSd	Interventricular Septum at End-Diastole
LVIDd	Left Ventricular Internal Dimension at Diastole
QTc	Corrected QT Interval

List of Tables

List of Figures

Publications

The research conducted in this thesis has resulted in publications in peer-reviewed conferences and presentations at international conferences. These outputs are listed below by category.

Journal Articles

- Ramchand S, Cole D, Xie X. Towards an interpretable early sepsis detection tool in index intensive care stay: machine learning algorithm development using electronic healthcare data from MIMIC-IV. Manuscript being redrafted for submission.

- Ramchand S, Ploszajski M, Cole D, Xie X. Explainable machine learning to uncover distinct phenotypic signatures of hypertrophic cardiomyopathy and Fabry disease. Under review in *BMC Medical Research Methodology*.Link

Conference Papers

- Ramchand S, Tsang G, Cole D, Xie X. RetainEXT: enhancing rare event detection and improving interpretability of health records using temporal neural networks. In: Proceedings of the *IEEE-EMBS International Conference on Biomedical and Health Informatics*. 2022. Presented at: BHI 2022; September 27–30, 2022; Ioannina, Greece. Link

- Ramchand S, Xie X. Augmenting infrequent relationships in clinical language models with graph-encoded hierarchical ontologies. In: Proceedings of the *International Conference on Artificial Intelligence in Healthcare*. 2024. Presented at: AIiH 2024; September 4–6, 2024; Swansea, UK. Link

Conference Abstracts

- Ramchand S, Cole D, Virdi A, Ploszajski M, Xie X. Using machine learning to distinguish Fabry disease from hypertrophic cardiomyopathy with ECG and ECHO data. *Molecular Genetics and Metabolism*. 2024;141(2):108014. Presented at: WORLD Symposium 2024; February 4–9, 2024; San Diego, USA. Link

- Ramchand S, Ploszajski M, Cole D, Xie X. Differentiating hypertrophic cardiomyopathy (HCM) and using AI. Presented at: *Cardiology Symposium Wales*; March 13, 2024; Cardiff, UK. Nominated for the Ian Williams Award.

- Ramchand S, Ploszajski M, Cole D, Xie X. Explainable machine learning to uncover distinct phenotypic signatures of hypertrophic cardiomyopathy and Fabry disease. Presented at: *British Inherited Metabolic Disease Group (BIMDG) Symposium*; June 11–12, 2024; Newport, Wales.

- Onongaya V, Ploszajski M, Ramchand S, Cole D, Xie X. Clinician feedback on the design features of an AI tool to identify Fabry disease. 2024. Presented at: AIiH 2024; September 4–6, 2024; Swansea, UK. Archived in Zenodo Link

Preamble

Living with a genetic disease like Fabry disrupts life in profound and often invisible ways. For some, it means a shortened life expectancy; for others, it means confronting daily obstacles, including loss of motor function, deteriorating memory, muscle atrophy, breathing difficulties, and progressive organ damage. The stories of patients like Cheryl and Bob highlight the human stakes behind rare diseases and the continued need for better tools, knowledge, and systems to support those affected.

Cheryl's journey reflects the immense physical, emotional, and social difficulties often faced by people with rare conditions. Growing up, she experienced symptoms such as burning pain in her hands and feet, frequent fevers, fatigue, and neurological issues. Despite this, she was initially diagnosed as a carrier at 16, reflecting that knowledge of Fabry disease manifestation in females was scarce at the time. It took years of suffering, numerous tests, and multiple consultations before the full impact of Fabry disease on her life was recognised. Cheryl also speaks to the mental toll and isolation that many patients experience, noting how difficult it can be for others to understand symptoms that are not always visible. She emphasises the value of patient communities, advocacy groups, and clinicians whose expertise and compassionate approach help patients feel heard, supported, and understood—and ultimately, move toward a diagnosis.

Bob's diagnostic journey was similarly complex. Over a decade of unexplained health issues and visits to 12 specialists across five NHS hospitals preceded his diagnosis in 2005. By then, Fabry disease had already caused significant damage to his kidneys and heart. His nephrologist explained that his kidney function had dropped to 15%, requiring urgent placement on the national donor list.

Bob's wife stepped forward as a donor. But the emotional and physical strain coincided with a sudden collapse in his remaining kidney function, leading to daily dialysis. Three months later, Bob received a kidney transplant from his wife. Yet complications followed:

his wife had CMV, a common but potentially serious virus, and despite preventive efforts, Bob contracted the infection shortly after the transplant. The resulting illness caused some of the most severe pain he had experienced.

Bob's life after diagnosis was marked by further complications—bleeding from port sites, mini-strokes, brain aneurysms—along with near-death experiences and ongoing anxiety about what might happen next. Managing a full-time job as a general manager, he pushed through the pain to provide for his family. Yet the disease took its toll. On one occasion, he experienced severe chest pain during an all-employee meeting and had to be rushed to the emergency room.

These stories reflect the reality that rare diseases are inherently difficult to diagnose and manage, often requiring the expertise of many to form the clinical picture. Fabry disease, like many other rare conditions, presents variably and often mimics more common disorders. Delays in diagnosis often result from overlapping symptoms, low prevalence, and the absence of clear early markers.

Throughout these journeys, the contributions of healthcare professionals—from general practitioners to nephrologists, cardiologists, geneticists, and transplant specialists—have been vital. Cheryl and Bob's care depended on the ongoing efforts of committed clinicians navigating diagnostic uncertainty, coordinating multi-specialist care, and responding to life-threatening complications. Their experiences underscore the importance of not only better education and tools for rare disease identification but also sustained support for the clinical teams who work tirelessly to care for these patients.

Approximately 1 in 17 people in the UK—over 3.5 million individuals—live with a rare disease [2]. In an already pressured system, the volume of patients means that access to specialised knowledge and coordinated care remains limited. Cheryl's and Bob's stories are powerful reminders of both patient resilience and clinician dedication. They call for greater investment in rare disease research, better integration of care, and more resources to ensure that no patient's condition goes unrecognised or untreated for lack of information.

In recognising the complexity of rare disease diagnosis and care, we acknowledge both the suffering that often accompanies these conditions and the ongoing commitment of those working to improve outcomes. Cheryl's and Bob's experiences are a call to action—not to replace clinical expertise, but to support it through better tools, more structured data, and integrated systems. This thesis responds to that call by exploring how computational approaches can help surface earlier signals, improve access to clinical information, and support more timely, informed decisions in the care of rare and under-represented conditions.

Chapter 1

Introduction

Contents

1.1 Motivation

1.1.1 Personalised Healthcare: A Transformative Vision

Imagine a healthcare system where every medical decision is tailored to a patient's genetic makeup and environment. This is the vision of personalised medicine, a paradigm shift that promises early, accurate diagnosis of rare diseases, prompt recognition of atypical presentations, and the prevention of adverse drug reactions. By adapting care to each patient's unique needs, personalised medicine can improve outcomes across the entire system.

The NHS faces increasing pressures due to an ageing population and rising healthcare complexity. Traditional models, focused on reactive responses to symptoms, are proving insuffi-

cient. The COVID-19 pandemic underscored these limitations, highlighting the importance of early detection and proactive management, especially for conditions that deviate from common clinical patterns [3].

1.1.2 The Cost of a Reactive Healthcare System

The prevailing healthcare model has long been characterised by a reactive approach [4–6], where interventions typically occur only after symptoms have developed or diseases have progressed. However, this approach is increasingly straining the healthcare system in an ageing society like the UK. For example, in Wales, the proportion of the population aged 65 and over rose from 18.4% in 2010 to 20.6% in 2020 [7]. This demographic shift is accompanied by higher rates of chronic diseases and long-term conditions such as cardiovascular disease, diabetes, and dementia [8], all of which require complex, ongoing care.

The limitations of reactive care are particularly apparent in chronic disease management. Conditions like type 2 diabetes often go undetected until patients present with complications [4], while cardiovascular interventions frequently begin only after significant events such as heart attacks, despite the existence of early warning signs [5]. Cancer care similarly reflects a reactive model, where interventions typically follow symptom onset rather than proactive, risk-based screening [6]. This delay in detection leads to poorer outcomes and increases treatment costs significantly. For example, delays in the early detection of heart failure cost the NHS £400 million for admissions [9].

The COVID-19 pandemic further exposed these vulnerabilities. Delays in treatment, for example, a four-week delay in cancer care, were associated with an increased risk of death of 6% to 13% [10]. These systemic shortcomings make clear the need for more preventive and flexible models of care.

1.1.3 Rare Diseases, Diagnostic Delays, and Data Imbalance

Conventional diagnostic models, which rely on rule-based reasoning and established clinical pathways, are effective for common conditions. These approaches rely on standard symptom clusters and predefined diagnostic routes, which work well when patients present with typical patterns. However, they face limitations when rare diseases or atypical symptoms deviate from these norms, often resulting in misdiagnosis or diagnostic delay.

Rare diseases frequently present with non-specific or atypical symptoms, making them difficult to diagnose using standard clinical pathways. The widely taught medical heuristic,

"When you hear hoofbeats, think horses, not zebras," illustrates a systemic bias toward common conditions [2], which can further complicate diagnosis.

The under-representation of rare diseases in research and clinical training further compounds these challenges. Small sample sizes and data sparsity limit the evidence base, hindering accurate and timely clinical decision-making [3, 11].

This diagnostic journey, known as the "diagnostic odyssey" [12], often results in long delays and uncertainty, which are reflected in clinical datasets as imbalanced, sparse, and temporally inconsistent records. These characteristics complicate the development of reliable machine learning models and highlight the broader issue of class imbalance, where rare or atypical conditions are systematically underrepresented in training datasets.

As a result, patients with rare diseases often face a prolonged path to diagnosis, involving multiple referrals, inconclusive tests, and frequent misdiagnoses. On average, it takes between 5.6 and 7 years to reach an accurate diagnosis [13]. This delay imposes significant physical, emotional, and financial burdens on patients and their families, postpones treatment, increases healthcare costs, and reduces the potential for positive health outcomes.

1.1.4 Moving Toward Proactive, Value-Based Health Care

In response to these challenges, Value-Based Healthcare (VBHC) has emerged as a forward-looking framework. Unlike volume-based models that reward service quantity, VBHC prioritises health outcomes, quality of life, and long-term care effectiveness. This approach aligns closely with the goals of personalised medicine, as it supports prevention, early diagnosis, and interventions that reflect patient-specific needs [14].

VBHC offers particular value in managing complex and rare conditions, where focusing on outcome metrics can help shift attention to underserved patient groups. It encourages a whole-system view that considers continuity of care, patient satisfaction, and sustainable cost management [11]. In Wales, national strategies such as "A Healthier Wales" and the "Rare Disease Implementation Plan" explicitly advocate for early diagnosis and integrated care models [15, 16].

1.1.5 The Role of Machine Learning

Machine Learning (ML) offers opportunities to support more proactive and personalised healthcare by enabling earlier and more accurate detection of disease patterns [17]. ML excels at analysing large-scale, high-dimensional data—including genomic information, imag-

ing, clinical records, and real-time sensor inputs—to identify patterns that may not be readily apparent to human observers.

Importantly, ML should not be viewed as a replacement for clinical expertise but as a complementary tool. These tools can identify rare or atypical presentations, assist in risk stratification, and provide decision support aligned with VBHC objectives. Integrating ML into clinical workflows enables a shift from reactive responses to anticipatory care pathways [3, 18].

1.1.6 Challenges of Modelling Imbalanced Clinical Data and Rare Events

Despite its promise, ML faces notable limitations when applied to clinical datasets, particularly those involving rare events. Data imbalance remains a core challenge: rare conditions are underrepresented in most training datasets, which can skew model performance and reduce its reliability [11]. Traditional algorithms often overfit to the majority class and fail to detect low-prevalence conditions.

Beyond class imbalance, rare events often exhibit high variability, making them harder to model using conventional techniques. Developing robust models requires methodological innovations such as synthetic oversampling, transfer learning, or uncertainty quantification. Moreover, metadata describing rarity, prevalence, and data provenance are essential for improving model transparency and interoperability.

These challenges highlight the need for domain-specific datasets and techniques that address the unique features of rare medical conditions. In particular, it is essential to consider imbalanced datasets, where statistical rarity, clinical heterogeneity, and low coding prevalence combine to create additional modelling challenges. Without careful design, there is a risk that data-driven tools could unintentionally reinforce diagnostic inequities instead of reducing them [19].

Integrating ML into clinical practice in a way that aligns with the principles of VBHC may help healthcare systems move toward providing more equitable and effective care. This thesis explores how data-driven methods can improve diagnostic accuracy and address the systemic and structural challenges that currently limit healthcare delivery for rare disease patients and others at the margins of existing models.

1.2 Contributions

The key contributions of our study can be summarised as follows:

Understanding the Patient Journey in the Digital Age

A comprehensive overview of how digital transformation reshapes the patient journey within the UK healthcare system. By examining the roles of primary, secondary, and tertiary care, this study highlights the complexities and delays in diagnosing rare diseases and atypical presentations, underscoring the need for data-driven approaches to support earlier and more accurate interventions.

Reviewing Data-Driven Methods for Rare Event Detection

A structured review of machine learning techniques for rare event detection in healthcare that clarifies the distinction between statistical imbalance, clinical rarity, and atypicality. It provides the conceptual foundation for the thesis's methodology and synthesises current limitations and opportunities in the field.

Addressing Data Imbalance and Missingness in Clithesis'sttings

A clinically informed preprocessing pipeline for intensive care datasets with moderate class imbalance and high missingness. Feature selection draws on clinician collaboration, while patterns of missingness are characterised to inform future data collection. By evaluating multiple deterioration endpoints, the study demonstrates how predictive features shift depending on the clinical score and identifies a novel physiological marker that may support earlier detection of subtle decline.

Enhancing Rare Event Detection through ML

Temporal representations and self-attention mechanisms are applied to model COVID-19 hospitalisation risk using longitudinal primary care data with pronounced class imbalance. The modelling approach demonstrates improved performance through event sequencing and timing. Feature attribution highlights the predictive value of underexplored clinical signals, such as nutrition referrals, offering insight into early deterioration in community settings.

Integrating Structured Medical Knowledge into Language Models

Structured clinical ontologies are embedded within pretrained language models using diverse embedding aggregation techniques. This method improves the representation of infrequently coded clinical conditions and supports low-data prediction tasks such as COVID-19 mortality and stroke classification from unstructured notes, while enhancing interpretability in clinical decision-making.

Building a Rare Disease Dataset for Diagnostic Research

A multimodal dataset is assembled containing cardiac test results from patients with Fabry disease and HCM, in partnership with hospital clinicians. This resource enables comparative studies of disease presentation, supports the development of data-driven diagnostic tools, and facilitates broader clinical research into rare cardiac conditions.

Differentiating Complex Cardiac Diseases with Similar Presentations

Machine learning models are developed to distinguish FD from HCM, two conditions with overlapping phenotypes but differing prevalence and underlying causes. By treating heterogeneity as a diagnostic challenge rather than a class imbalance issue, the analysis highlights the role of data-driven tools in supporting differential diagnosis where standard pathways may fail.

Augmenting Predictive Models with Language-Based Explanations

A prototype risk calculator pairs machine learning predictions with natural language explanations generated by large language models. The interface enables interactive inspection of patient-specific risk factors and highlights variables most influential to each prediction. Prompt engineering methods refine the clarity and clinical usefulness of explanations. Although only one round of clinician feedback was completed, the prototype demonstrates how interpretability techniques align model logic with practitioner reasoning.

These contributions span data handling, method development, and clinical application. Collectively, they aim to advance more equitable and proactive healthcare through data-driven tools that are technically rigorous, clinically relevant, and practically implementable.

1.3 Thesis Overview

This thesis explores the critical intersection of data-driven approaches and healthcare, focusing on detecting and managing rare events. In an era where big data and Artificial Intelligence (AI) are transforming numerous sectors, healthcare stands to benefit significantly from these advancements. However, the unique challenges of imbalanced clinical scenarios, whether due to heterogeneity or rarity, require innovative solutions beyond conventional data analysis techniques.

The remainder of Chapter 1 outlines the document structure and its contributions to the wider knowledge base. The thesis is structured into several chapters, each addressing a specific facet of this complex problem:

Chapter 2, "Anomalies Amidst a Sea of Data", examines the role of ML and Deep Learning (DL) in rare event detection. It discusses modelling techniques, the challenges of predicting low-frequency events, strategies for addressing class imbalance, and the latest developments in the field. This chapter sets the stage for understanding the broader problem space and the need for advanced analytical approaches, contributing to more robust methods for identifying rare conditions and atypical presentations.

Chapter 3, "Data-Driven Approaches for Rare Event Detection in Healthcare", explores essential preprocessing and feature engineering processes for building"effective ML models. It addresses the challenges posed b" healthcare data, including heterogeneity, high dimensionality, and unstructured formats. By evaluating techniques for data cleaning, feature selection, dimensionality reduction, and the integration of domain knowledge, this chapter highlights the importance of data readiness. The use of medical ontologies further supports the interpretability and effectiveness of predictive models.

Chapter 4, "Navigating Imbalance and Missing Data in Clinical Settings", focuses on the case of sepsis—a condition often challenging to detect due to its s"btle onset. It explores the limitations of existing diagno"tic scores and the challenges posed by missingness in Intensive Care Unit (ICU) data, presenting a clinically informed approach to data preprocessing. This chapter contributes to developing more reliable and actionable predictive models that better handle real-world imbalance and data quality issues.

Chapter 5, "Mining Temporal Patterns in a Pandemic", shifts attention to COVID-19 and the use of primary care data to assess hospitalisation risk. It introduces temporal modelling using self-attention and examines opportunities to capture risk progression from community-level records. This chapter underscores the role of temporal pattern mining in anticipating patient deterioration and contributes strategies for proactive care during public health crises.

Chapter 6, "Ontological Osmosis: Infusing Structure into Language Understanding", explores how structured clinical knowledge can be embedded in unstructured" text representations. It examines the use of language models and graph neural networks and proposes a method for integrating medical ontologies into deep learning frameworks. This chapter contributes techniques for improving model coherence and performance, especially in low-data or sparsely coded conditions.

Chapter 7, "Decoding Rarity: ML to Distinguish Complex Cardiac Diseases with Similar Presentations", investigates how machine learning can aid the differential diagnosis of FD and HCM—two phenotypically overlapping cardiac conditions. It discusses the development of a bespoke dataset and evaluates the performance of models trained on multimodal cardiac test data. This chapter offers insights into the diagnostic challenges of disease heterogeneity and explores data-driven tools to support clinical reasoning.

Chapter 9, "Conclusion and Future Work", summarises the thesis's main contributions and suggests future directions. It reflects on the work's practical, technical, and clinical implications to enhance the design and deployment of data-driven systems for rare event detection.

Together, these chapters advance the design, development, and evaluation of methods for identifying rare, underrepresented, and atypical clinical events in imbalanced healthcare datasets. The thesis contributes to an emerging vision of healthcare that is more personalised, proactive, and equitable—one in which data-driven tools can assist clinicians in delivering timely and informed care, even in the face of uncertainty or rarity.

Chapter 2

The Patient Journey in the Digital Age

Contents

2.1 The Digital Transformation of the NHS

The United Kingdoms (UKs) NHS is one of the largest publicly funded healthcare systems globally. Established in 1948, it provides universal healthcare free at the point of use through a tiered structure comprising primary, secondary, and tertiary care services [20]. This framework ensures patients receive appropriate care, from routine consultations to complex specialist interventions.

In recent years, the NHS has undergone a digital transformation, adopting Electronic Health Records (EHRs), telemedicine, online appointment systems, and algorithmic diagnostic

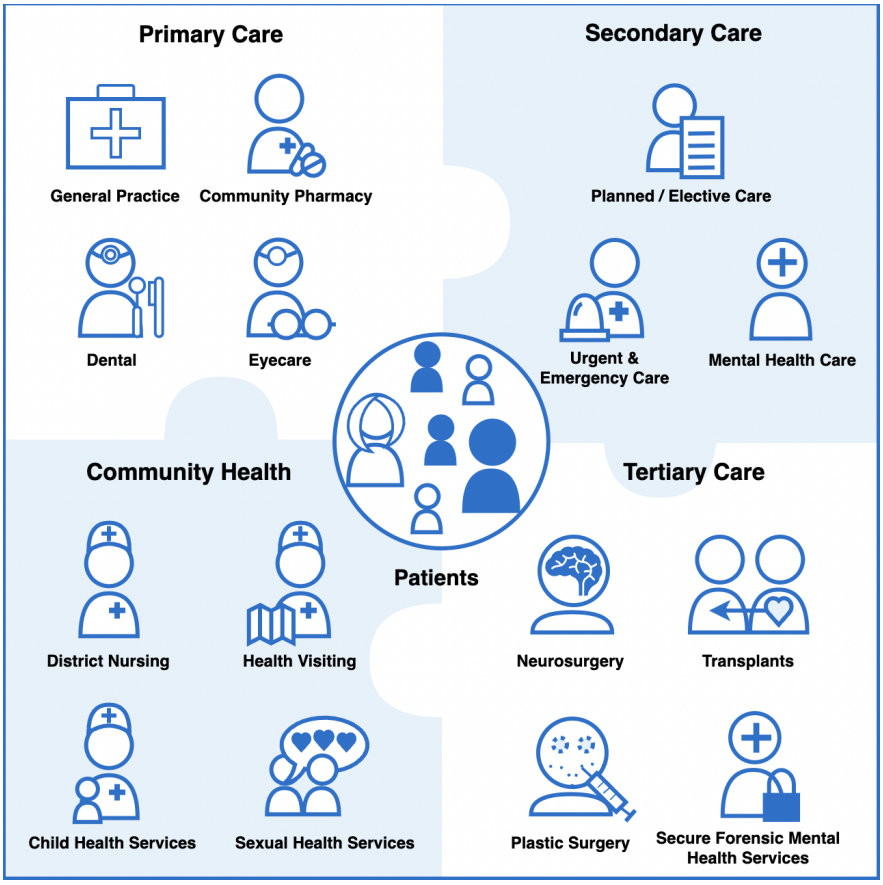


Figure 2.1: Structure of the NHS healthcare system [1]

tools [21]. These innovations offer significant efficiencies and expand access to care. As the volume of patient data grows, so does the potential to research and understand the evolving health needs of the population [22].

Yet despite the ever-increasing volume of data, data on rare diseases remains limited compared to other common ailments. Sporadic data on rare diseases highlight the challenges of data fragmentation [23]. Diagnosing these conditions is often hindered by sparse structured data and inherent diagnostic uncertainty. Moreover, clinical knowledge is recorded across independent data sources like unstructured notes, diagnostic codes, laboratory results, and fragmented care records, making it difficult for clinicians to access the full spectrum of information necessary for accurate diagnoses [24]. Furthermore, inconsistent coding systems and a lack of standardisation across different stages of care exacerbate these issues.

This fragmentation often results in delays and misdiagnoses, particularly for rare diseases,

where the absence of clear diagnostic criteria can lead to prolonged diagnostic odysseys [12]. While AI and ML hold promise for identifying rare diagnostic patterns to serve as a clinical decision support tool, these technologies must address existing data integration and interpretation limitations to be genuinely effective.

This chapter explores the digitalisation of the NHS, examining how data fragmentation, inconsistent coding systems, and imbalanced clinical data create barriers to accurate diagnosis and timely care. It sets the stage for subsequent discussions on how data-driven approaches can improve the diagnosis and management of rare and atypical conditions.

2.2 Patient Data Types

This section examines how data is collected throughout a patient's journey in the NHS. This is especially pertinent as our research draws on multiple data sources—both structured and unstructured—to develop robust methods for distinguishing between rare diseases, such as FD and HCM [25], and for early detection of conditions that present heterogeneously, such as sepsis. Understanding the data sources available, their potential biases and limitations, and data quality is vital to determine appropriate methods for preprocessing the data before applying AI techniques.

When individuals seek care within the UKs NHS, they enter a system that is at once comprehensive in its ambition and constrained in its delivery. Patient journeys typically begin in primary care, with progression to secondary or tertiary services depending on the complexity, persistence, or severity of the condition [26].

Patient data is recorded digitally as EHRs and is typically segmented into structured (e.g., coded diagnoses), semi-structured (e.g., tick-box consultation templates), and unstructured (e.g., free-text notes) forms. Unstructured data sources provide valuable insights often absent from structured datasets, offering opportunities to enhance existing knowledge bases and enrich data for more comprehensive health research [27]. For instance, open-ended doctors' notes from the initial patient assessment can supplement longitudinal laboratory, clinical, and survey data. The following sections detail the different data gathered in each level of care.

2.2.1 Primary Care

Structured fields (e.g., SNOMED CT, Read Codes) are recorded for billing and audit purposes. However, the more nuanced aspects—symptom evolution, patient intuition and social

context—are often confined to free-text notes or omitted entirely. This results in data that are fit for billing but may not fully reflect patients' health status. General Practices (GPs) may hesitate to use specific codes when unsure of the diagnosis [28], which is often the case in the early stages of rare diseases. As a result, the early stages of atypical conditions are underrepresented in the patient's digital footprint. Therefore, estimates suggest that over 70% of diagnostic detail is embedded in unstructured notes [29]. However, this content is difficult to extract or analyse, especially across care boundaries. For patients with rare diseases, the critical early signals may be recorded in inaccessible formats or not at all, skewing both individual care and population-level insights.

2.2.2 Secondary Care

Secondary care settings, including general hospitals and specialist clinics, collect a wide range of data during patient interactions, investigations, and treatments. This data is crucial for diagnosis, treatment planning, and ongoing patient management for common and moderately complex conditions.

Structured Data

- Patient demographics and admission details
- Vital signs and regular observations (e.g., blood pressure, heart rate, temperature)
- Standard laboratory test results (e.g., complete blood count, liver function tests, electrolytes)
- Medication administration records
- Diagnostic codes for common and moderately complex conditions
- Procedure codes for routine interventions
- Appointment and attendance records
- Length of stay for inpatient admissions

Semi-structured Data

- Clinical assessment forms for each speciality (e.g., Cardiology, Gastroenterology)

- Nursing care plans and daily notes
- Discharge summaries following standardised templates
- Patient-reported outcome measures (PROMs) for common conditions

Unstructured Data

- Clinical notes from specialist consultations
- Surgical notes for common procedures
- Radiology reports for standard imaging (e.g., X-ray, CT, basic MRI)
- Pathology reports for routine biopsies
- Patient history and symptom descriptions

Qualitative Data

- Specialist's observations on patient progress for common treatments
- Notes on patient's quality of life related to common conditions
- Basic psychological assessments
- Patient's reported experiences with standard treatments

Imaging and Signal Data

- Standard radiological images (X-rays, basic CT scans)
- Common ultrasound images and videos (e.g., abdominal, cardiac)
- Basic electrocardiogram (ECG) recordings

Specialised Test Results

- Common endoscopy findings
- Standard biopsy results
- Routine microbiology and culture results

- Basic immunology test results (e.g., autoantibody screens)

The range of data collected in secondary care provides a more comprehensive picture of a patient's condition, particularly for common and moderately complex cases. It enables specialists to make informed decisions about diagnosis and treatment for a wide range of conditions that do not require highly specialised tertiary care interventions.

2.2.3 Tertiary Care

As the pinnacle of medical specialisation, tertiary care centres collect highly specialised and complex data often unavailable in primary or secondary care settings. This data is crucial for managing rare diseases, complex conditions, and cases that require cutting-edge treatments or technologies.

Structured Data

- Advanced genetic test results (e.g., whole genome sequencing, exome sequencing)
- Specialised biomarker measurements for rare diseases
- Detailed immunological profiles
- Advanced neurophysiological measurements (e.g., complex EEG patterns)
- Specialised organ function tests (e.g., advanced cardiac output measurements)
- Clinical trial enrolment and protocol adherence data
- Rare disease registries data

Semi-structured Data

- Multidisciplinary team meeting records for complex cases
- Specialised assessment forms for rare conditions
- Detailed treatment protocols for experimental therapies
- Long-term follow-up plans for complex chronic conditions
- Research study participation records

Unstructured Data

- Detailed clinical notes from sub-specialist consultations
- Complex surgical notes for highly specialised procedures
- Advanced imaging reports (e.g., functional MRI, PET-CT fusion imaging)
- Molecular pathology reports for rare cancers or genetic disorders
- Case reports for novel or unusual presentations of diseases

Qualitative Data

- Patient-reported outcomes for experimental treatments
- Quality of life assessments for rare or complex conditions
- Psychological impact evaluations for novel therapies
- Family genetic counseling session notes
- Patient experiences with cutting-edge medical devices or prosthetics

Imaging and Signal Data

- Advanced neuroimaging (e.g., tractography, perfusion imaging)
- Molecular imaging data (e.g., PET scans with novel tracers)
- High-resolution microscopy images for rare cellular abnormalities
- Complex cardiac imaging (e.g., 4D flow MRI)
- Intraoperative imaging from robotic or computer-assisted surgeries

Specialised Test Results

- Results from novel diagnostic tests for rare diseases
- Pharmacogenomic profiling data
- Specialised tissue and organ function tests

- Advanced immunotherapy response measurements
- Metabolomic and proteomic analysis results

This highly specialised data collected in tertiary care settings provides unprecedented insights into complex and rare conditions, enabling personalised treatment plans and advancing medical knowledge. Managing this data presents unique challenges, such as the need for sophisticated integration systems, ensuring interoperability, and maintaining privacy for sensitive genetic and molecular data. Tertiary care data often contributes to national and international research collaborations, advancing global understanding of rare diseases and complex conditions. This data-driven approach is crucial for pushing the boundaries of medical science and offering hope to patients with the most challenging health conditions.

2.2.4 Fragmentation Across Levels of Care

In the NHS, patient data is scattered across different care settings. Primary care records typically focus on patient history and initial presentations, while secondary care adds more detailed investigations and diagnoses. Tertiary care, on the other hand, contributes advanced diagnostics, specialist insights, and genetic data. However, these records are often stored in separate systems that don't communicate well with each other.

Fragmentation of data creates persistent silos within the system. For instance, a clinician may receive a genetic diagnosis in tertiary care, but it may never be added to the patient's primary care record. Discharge summaries also tend to oversimplify complex care pathways. A multi-year journey through rheumatology, neurology, and genetics, for example, might be condensed into something like "patient under specialist review," which masks the complexity and difficulty of the diagnostic process. When records don't flow seamlessly across the system, neither does the knowledge, and it's often up to individual clinicians—or the patient—to try to piece it all together.

2.2.5 Data Imbalance in the Patient Journey

As patients move through different tiers of care, several types of data imbalances accumulate, shaping diagnostic outcomes:

- **Class imbalance:** Rare diseases are statistically under-represented in EHR datasets, skewing diagnostic algorithms towards more common conditions [4, 11].

- **Feature imbalance:** Key features—such as genetic markers or specialist assessments—are typically confined to tertiary care and may be absent from earlier records [13].
- **Temporal imbalance:** Data clusters around acute episodes or hospital visits; early, subtle symptom data is sparse or missing [11].
- **Contextual imbalance:** Social and behavioural data, often crucial for diagnosis and management, are inconsistently captured and rarely structured [30].
- **Volume imbalance:** Tertiary care centres, which generate some of the richest diagnostic data, deal with fewer patients overall, limiting the data available for training robust AI systems [3].
- **Format imbalance:** High-value data is often stored in static documents, PDFs, or non-standardised formats that hinder integration into larger datasets or decision-support tools [11].
- **Overlap-Induced Diagnostic Ambiguity:** Conditions with highly similar phenotypes (e.g., a rare disease and a more common condition) occupy overlapping regions in the feature space, causing diagnostic ambiguity. This overlap leads to challenges in distinguishing between conditions, despite their clinical significance. Unlike class imbalance, this issue arises not from data frequency but from the similarity in symptom profiles, which blurs the decision boundaries of diagnostic algorithms [31, 32].

These imbalances are cumulative and systemic, reflecting broader issues in documentation practices, time constraints, and system design. Their impact is especially pronounced for patients who fall outside typical diagnostic trajectories—those with rare, multi-system, or socially complex conditions.

2.3 The Diagnostic Odyssey in Rare Diseases

For many patients, particularly those with rare, complex, or multi-system conditions, the journey to a correct diagnosis is long, uncertain, and emotionally taxing [2]. With thousands of rare diseases, each affecting fewer than 1 in 2,000 people in the UK, it is unrealistic to expect GPs to be familiar with all these conditions and the diverse ways they may present [2].

When a rare disease is suspected, securing the appropriate referral becomes a diagnostic hurdle in itself [12]. Referral pathways are structured around explicit criteria to manage specialist demand. However, rare diseases often do not conform to these early-stage criteria, and referrals may be delayed or rejected.

When patients eventually enter secondary or tertiary care, they often encounter further challenges. Rare diseases frequently span organ systems and medical domains, necessitating multiple specialist referrals across hospital departments or healthcare trusts [12].

Managing these complex care pathways requires both clinical insight and substantial administrative effort. For example, a patient referred to multiple clinics may discover that each department maintains separate electronic health records with limited access to prior consultations or test results. As a result, clinicians may unknowingly duplicate investigations or miss critical longitudinal clues—delays that frustrate patients and can directly impact care quality and safety, such as through redundant imaging, medication conflicts, or missed diagnostic connections [33].

2.3.1 Compounding Challenges and Their Effects on Data Imbalance

For patients with rare diseases, this filtering effect is even more pronounced. Their symptoms often defy tidy categorisation, emerge gradually, or interact across body systems. Early encounters may be under-recorded, mistranslated into common diagnoses, or left as vague notes that do not meet referral thresholds. Without clear codes or integrated longitudinal records, these patients risk becoming invisible within their records. Beyond delays and fragmentation, rare disease patients face additional structural and lived challenges that further entrench diagnostic inequity and contribute to long-standing data imbalances:

Patient-Led Coordination: In the absence of integrated care systems, many patients become their case managers—tracking appointments, test results, and care plans while facilitating communication between multiple specialists. These interactions often go undocumented in formal records, creating blind spots in the healthcare data landscape that hinder care coordination, obscure symptom evolution, and complicate longitudinal case analysis. The lack of visibility into these informal exchanges and patient-led insights can result in missed diagnostic opportunities, fragmented treatment plans, and safety risks due to untracked medication changes or duplicated investigations.

Polypharmacy and Medication Complexity: Managing a rare disease often involves multiple medications prescribed across specialities. This increases the risk of drug interactions

and places a greater burden on patients to maintain a coherent treatment plan. The nuances of these regimens—timing, side effects, off-label uses—are rarely recorded in full, leading to underestimation of their impact on clinical care and data analysis.

Barriers to Diagnosis and Treatment Eligibility: Diagnostic pathways often include strict approval thresholds tied to treatment subsidies or care eligibility. Access may be delayed or denied for patients whose presentations do not meet these thresholds, whether due to early-stage symptoms, non-conforming test results, or atypical demographic markers. The high cost of many orphan drugs further intensifies this gatekeeping, creating disincentives for formal diagnosis unless necessary.

Invisibilised Histories and Demographic Skew: Many rare diseases have historically been diagnosed in populations with greater access to specialist care. This has led to diagnostic archetypes that reflect narrow demographic norms, potentially obscuring how rare diseases present in underrepresented groups, by race, gender, age, or geography. As a result, diagnostic algorithms and training materials may be biased toward majority-population data, perpetuating underdiagnosis in minority communities.

Together, these challenges compound rarity with complexity, reinforcing the systemic limitations outlined earlier in the NHS's structural design. The fragmented flow of information, rigid referral criteria, and inconsistencies in documentation become even more consequential in the context of rare diseases. These factors delay individual diagnoses and impede the generation of reliable population-level data, limiting the healthcare system's capacity to learn from and respond to the needs of rare disease patients. This underscores the urgency of designing data systems, clinical workflows, and governance models that account for—and actively address—this compounded complexity. In a data-driven era, such exclusions are not just clinical—they shape the evidence base for future diagnoses.

2.4 Summary

The challenges patients face while navigating the NHS, particularly those with rare, complex, or atypical conditions, are systemic but not immutable. These include rigid referral thresholds, fragmented records, inconsistent coding practices, and narrow diagnostic archetypes, often excluding patients who do not conform to expected clinical trajectories. For rare disease patients, these structural limitations are compounded by data invisibility, demographic bias, and the burden of self-managed care.

Addressing these challenges requires more than merely upgrading digital infrastructure. It demands a deliberate focus on improving data quality, interpretability, and integration throughout the patient journey. This includes acknowledging and preserving early, ambiguous, or incomplete clinical signals, addressing underrepresentation in coding systems, and better using unstructured notes and under-coded conditions. The research that follows focuses on precisely these fronts: identifying clinical patterns within imbalanced and temporally sparse data, developing self-attention methods to detect overlooked presentations, and exploring how intelligent risk calculators and Large Language Models (LLMs) can work in tandem to enhance interpretability, contextualisation, and diagnostic support.

By building systems that recognise complexity rather than filter it out, we can begin to redress diagnostic inequities and support more timely, accurate, and inclusive care. This work lays the foundation for developing clinical tools and data governance frameworks that are technically robust and socially and ethically responsive to the realities of rare disease care.

Chapter 3

Data Imbalance and Rare Event Detection in Healthcare

Contents

3.1 Introduction

Building upon the previous chapter's exploration of the UKs NHS and the challenges of diagnosing rare and complex conditions, this chapter shifts focus to the computational challenges posed by data imbalances in healthcare. Specifically, it investigates how imbalances—resulting from low-prevalence events like rare diseases, adverse drug reactions, or emerging infectious outbreaks—complicate predictive modelling and clinical decision-making [4, 11].

Rare events in healthcare are typically defined not only by their infrequency but also by their clinical impact [34]. Rare diseases or adverse drug reactions deviate from typical clinical patterns, making early identification difficult. The scarcity of data for rare events leads to imbalanced datasets, where the minority class (e.g., rare diseases) is underrepresented compared to the majority class (e.g., non-rare disease individuals). Although class imbalance is the most commonly studied form, imbalances also arise due to gaps in feature availability, temporal inconsistencies, and incomplete or inconsistent documentation across care levels [34]. These issues skew model predictions and increase the risk of misdiagnosis and delayed intervention. Addressing these imbalances is crucial to improving the detection of rare events, as timely and accurate identification can enable earlier interventions, targeted care, and better patient outcomes [13, 34].

One of the primary challenges in healthcare is the diagnosis of rare diseases. Rare diseases collectively affect approximately 3.5% to 5.9% of the global population, translating to over 260 million people worldwide [2]. Despite the individual rarity of these diseases, the cumulative impact is profound, with over 7,000 distinct rare diseases identified [2]. However, the diagnostic process for these conditions is often complex and resource-intensive, with significant delays in diagnosis due to symptom overlap with more common diseases and limited clinician expertise [2].

In this context, computational methods, particularly those grounded in ML, offer significant promise for overcoming the challenges of data imbalance. When developed responsibly

and integrated into clinical workflows, these techniques can augment human expertise, synthesise high-dimensional data, and uncover patterns that may go unnoticed. However, applying these methods to rare event detection requires addressing challenges that are not as pronounced in more balanced datasets. In its various forms, imbalance compounds the difficulty of detecting rare events, affecting the model's generalisability and diminishing predictive accuracy. Challenges such as high dimensionality, data heterogeneity, temporal inconsistencies, and the ambiguity between what constitutes a rare event and what is an anomalous case (especially when early symptoms overlap with more common conditions) further complicate detection efforts [3, 11].

Thus, the interplay between data imbalance and rare events necessitates a nuanced approach to predictive modelling. Rare events often create imbalanced datasets by class and feature availability, historical data, and the lack of contextual data. This chapter presents a comprehensive overview of rare event detection, framed within the context of imbalanced datasets. It begins by characterising the nature of rare event data and discussing the various sources of imbalance that affect model performance. The chapter then reviews the strategies employed to mitigate these imbalances, including data-level techniques (e.g., oversampling and undersampling), algorithmic approaches (e.g., cost-sensitive learning), and ensemble methods (e.g., combining multiple models to improve prediction accuracy). The chapter also highlights the importance of using specialised evaluation metrics for imbalanced data, as traditional metrics may provide misleading conclusions in such contexts [35].

The chapter further examines the importance of interpretability in ML models, especially in healthcare, where decision-making based on model predictions can have significant consequences. Clinician trust is critical, and understanding the reasoning behind a model's predictions is key to ensuring its acceptance in clinical practice [19]. The chapter concludes by identifying research gaps and emerging directions in rare event detection, setting the stage for the following chapters, which delve into case studies of rare disease prediction using NHS-linked data.

By framing the computational challenges of rare event detection within the broader context of data imbalance, this chapter builds upon the patient-centric and systemic focus introduced earlier. It is a conceptual and methodological reference for the data-driven approaches explored throughout our work.

3.2 Rarity Levels in Healthcare Data

Understanding the varying rarity levels in healthcare data is essential for developing effective rare event detection strategies. Rarity is not a binary concept; it exists on a spectrum, with events ranging from those that occur rarely to those that are still uncommon but occur more frequently. Recognising these levels allows for tailored detection methods that better address the challenges posed by each category.

Rarity in healthcare events can be classified into four levels, based on the work of Shyalika *et al.* (2024) [36]:

- **R1 - Extremely Rare (0-1%):** These events are exceedingly rare, accounting for 0-1% of occurrences in the dataset. In healthcare, examples include extremely rare genetic disorders or adverse reactions to treatments that affect only a small fraction of the population.

- **R2 - Very Rare (1-5%):** Events in this category occur more frequently than R1 events but still represent a small proportion of the dataset (1-5%). These may include rare cancers or certain orphan diseases that, although uncommon, require specialised detection methods.

- **R3 - Moderately Rare (5-10%):** These rare events (5-10%) are more common than R1 and R2 but still infrequent. They may include certain autoimmune diseases or genetic syndromes affecting a larger portion of the population.

- **R4 - Frequently Rare (10+%):** While considered rare in some contexts, these events occur more frequently than other rare conditions (more than 10%). Examples might include specific mental health disorders or less rare genetic mutations.

This classification of rarity helps guide the selection of rare event detection methods. By aligning detection techniques with the event's frequency, the approach can be better suited to handle the specific challenges posed by each rarity category, ultimately improving diagnostic accuracy and patient outcomes.

3.3 Understanding Rare Event Data: Characteristics and Challenges

Rare events manifest across various domains, each with unique characteristics that necessitate specialised detection strategies. In fields like finance, events such as the "Black Monday" crash of 1987 illustrate how infrequent occurrences can trigger global crises, leading to widespread economic impacts [37]. Similarly, the insurance industry faces challenges from rare but catastrophic natural disasters such as hurricanes and volcanic eruptions [38].

In healthcare, rare events similarly carry significant implications. The emergence of novel pathogens like HIV/AIDS highlights how rare events can escalate rapidly, turning from isolated occurrences into global health crises [39]. Similarly, rare diseases—by their very definition—pose complex diagnostic challenges, often sharing symptoms with more common conditions and leading to diagnostic delays or misdiagnoses [40].

These examples underscore the multifaceted nature of rare events, where rarity is often compounded by complexity. Effective rare event detection requires understanding the data's inherent characteristics and the challenges in accurately capturing those events.

Particularly for rare diseases, the need for vigilance is heightened. Despite affecting over 300 million people globally, rare diseases often go undiagnosed or misdiagnosed due to symptom overlap with more common diseases, compounded by the limited data available for these conditions [2, 41]. This scarcity of data adds an additional layer of difficulty, demanding more sophisticated techniques for accurate prediction and identification.

3.3.1 Class Imbalance and Rare Events

At the core of rare event detection lies the challenge of class imbalance, which arises directly from the event's rarity. This imbalance occurs when the majority class (e.g., common diseases) vastly outnumbers the minority class (e.g., rare diseases), skewing predictive models. These models develop a bias toward the majority class, resulting in poor sensitivity for the minority class and delaying the detection of rare events.

A clear example of this is the diagnosis of Hutchinson-Gilford Progeria Syndrome (HGPS), an extremely rare genetic disorder with an incidence of about 1 in 18 million live births [42]. Developing accurate diagnostic models for such rare conditions is inherently difficult due to the scarcity of positive instances. Similarly, detecting rare adverse drug reactions (ADRs) amidst the majority of non-adverse events introduces another layer of class imbalance [43].

Advanced techniques, such as oversampling (e.g., SMOTE), undersampling, and synthetic data generation, help mitigate this imbalance by adjusting the dataset distribution [44, 45]. Moreover, cost-sensitive learning and anomaly detection algorithms prioritise rare events during model training, ensuring they receive sufficient attention despite the class imbalance [46].

3.3.2 High Dimensionality and Heterogeneity

The challenges of high dimensionality and heterogeneity in rare event data further complicate detection efforts. In healthcare, particularly in rare diseases, datasets often contain thousands of variables spanning clinical notes, lab results, and imaging studies. However, only a small subset of these features may be relevant to identifying a particular rare disease, leading to an excess of irrelevant or redundant data that introduces noise into the model.

Additionally, the heterogeneity of rare disease symptoms adds another layer of complexity. Conditions like Behçet's disease, a multi-systemic disorder that presents with a wide range of symptoms across different organ systems, make developing a unified diagnostic model more difficult [47].

The diversity of data sources—ranging from EHR and medical imaging to wearable devices—further complicates data integration. Each source has its format, scale, and reliability, which challenge creating a consistent, unified model.

To address these challenges, researchers use dimensionality reduction techniques, such as Principal Component Analysis (PCA) and autoencoders, to reduce the high-dimensional data while retaining key features. Feature selection methods like LASSO regression and random forests help identify the most relevant variables. Multimodal data fusion and transfer learning techniques combine information from diverse sources, enhancing model performance.

3.3.3 Temporal Properties

Another challenge in rare event data is the presence of complex temporal patterns. Some rare diseases exhibit seasonal variations, and the incidence of certain conditions may change over long-term trends, requiring models that can adapt to non-stationary data. Epidemic outbreaks or rare events may also occur in bursts or clusters, requiring spatial-temporal analysis and cluster detection algorithms.

For example, Kawasaki disease, a rare paediatric vasculitis, shows seasonal patterns and occasional outbreaks, which must be captured in diagnostic models [48]. Similarly, rare in-

fectious diseases such as the Nipah virus emerge in clusters, and early detection relies on identifying these unusual patterns in space and time [49].

To model these temporal dependencies, time series analysis techniques such as ARIMA [50], recurrent neural networks [51], and transformer-based models [52] are increasingly being used. These advanced methods can account for exogenous variables and long-term dependencies, critical for rare event data analysis.

3.3.4 Importance of Metadata

The importance of metadata compounds the challenges posed by class imbalance, high dimensionality, and temporal complexity. In healthcare, metadata such as patient demographics, medical history, and environmental exposures can offer crucial insights for detecting rare events. Researchers can enhance diagnostic accuracy by integrating metadata with clinical data through multimodal fusion techniques, especially for rare diseases.

In rare genetic disorders, metadata such as family history, parental consanguinity, ethnic background, and geographical location can provide crucial clues for diagnosing rare diseases. Certain conditions, for instance, are more prevalent in specific ethnic groups, and knowing this can increase the probability of early detection [53, 54].

Effective metadata integration requires expertise in domain knowledge and advanced data integration techniques. Knowledge graphs, hierarchical modelling, and transfer learning are essential for incorporating this contextual information into predictive models [55].

3.3.5 The Broader Perspective

Context is paramount in rare event analysis, and the evolving healthcare landscape further impacts the detection and prediction of rare events. Factors such as changes in diagnostic criteria, increased public and clinical awareness, and the introduction of new screening programmes can alter how frequently rare events are observed or recorded.

For example, expanding newborn screening programmes has significantly increased the detection of rare metabolic disorders, largely due to the incorporation of Next-generation sequencing (NGS). Technology advancements have enabled earlier detection and interventions, ultimately improving patient outcomes [56]. Similarly, recognising rare autoinflammatory syndromes has increased due to international collaboration and improved diagnostic criteria [57].

Incorporating contextual factors such as societal changes, environmental influences, and diagnostic shifts requires interdisciplinary approaches that combine data science with expertise

from fields like epidemiology, environmental science, and sociology. This broader perspective is crucial for developing effective strategies to detect, predict, and manage rare events in healthcare.

The challenges associated with rare event data, compounded by data imbalances, highlight the need to address data quality, accuracy, and validation issues. Misclassification of rare events can significantly skew analysis results, underscoring the necessity for robust validation and verification processes. Data quality initiatives, such as expert review and cross-validation, are essential for ensuring the integrity and reliability of rare event detection models.

In the healthcare domain, rare event data characteristics present challenges and opportunities. The class imbalance in rare disease diagnosis, the high dimensionality of clinical data, and the complex temporal patterns of disease progression all contribute to the difficulty of detecting and managing rare health events. For instance, in clinical trials for rare diseases, ensuring data accuracy is critical due to small sample sizes, and methods such as bootstrapping are often employed to assess the stability of findings [58].

In summary, challenges from class imbalance, high dimensionality, temporal complexity, and contextual factors create a dynamic landscape for rare event detection. Addressing these challenges through innovative data processing techniques, better analysis methodologies, and improved data quality is essential for advancing our ability to detect, predict, and manage rare events in healthcare and beyond.

3.4 Data Processing for Rare Event Detection

Improving the performance of rare event detection models, especially with high-dimensional data, relies heavily on effective feature engineering and selection. These processes focus on selecting meaningful features and identifying the most relevant ones, enhancing model efficiency and interpretability. According to Guyon and Elisseeff (2003), feature selection reduces dimensionality, alleviates the curse of dimensionality, and improves model performance—critical factors in rare event detection [59].

3.4.1 Data Cleaning

Data cleaning is critical in preparing datasets for predicting rare clinical events. The rarity and complexity of such events often mean that the available data may be sparse, imbalanced, and

contain errors or inconsistencies. High-quality, reliable data are essential to developing robust models that predict these uncommon occurrences accurately.

In rare clinical events, data sifting and filtering refine large volumes of clinical data to identify the most relevant information. This process involves removing duplicate records, excluding irrelevant data entries, and selecting cases that meet specific inclusion criteria. For instance, Liu *et al.* (2018) emphasised the importance of filtering out irrelevant records to focus on significant predictors of postoperative complications in colorectal surgery [60].

While imputation is a common strategy for handling missing data, it is essential to note that over-reliance on simplistic imputation methods (e.g., mean imputation) can introduce bias into the data, especially when dealing with rare events. More sophisticated techniques, such as MICE or k-NN imputation, can provide better estimations, considering the relationships between variables and improving model performance.

Handling missing data through imputation is crucial, as rare events often have incomplete datasets due to their infrequency. Simple imputation methods, such as replacing missing values with the mean or median, may not be adequate. Advanced imputation techniques like Multiple Imputation by Chained Equations (MICE) or k-Nearest Neighbours (k-NN) imputation provide more accurate estimations by considering the relationships between variables [61]. For example, Rafsunjani *et al.* (2019) utilised MICE to address missing values when predicting failures in clinical equipment [62].

Noise removal is critical to eliminate erroneous or irrelevant data that could obscure important patterns associated with rare clinical events. Techniques such as the Edited Nearest Neighbours (ENN) method help remove noise by eliminating instances that differ significantly from their neighbours. Ashraf *et al.* (2023) applied ENN to improve the detection of wrong-way driving incidents—a rare but critical event—by cleaning the dataset of noisy entries [63].

Text-based cleaning techniques are vital for textual data, particularly in clinical narratives and patient records. Processes such as converting text to lowercase, removing punctuation, stop words, and numerical values, and stemming and lemmatisation help standardise the text data. Wong (2016) demonstrated the use of these techniques to preprocess clinical text data when detecting rare medication errors [64].

Similarly, when working with imaging data used in diagnosing rare conditions, image processing techniques are necessary to improve image quality and enhance the performance of DL models. Salvi *et al.* (2021) provided a comprehensive review on the impact of pre- and post-image processing techniques within DL frameworks for digital pathology image analysis [65].

Their work underscores the importance of applying preprocessing methods like normalisation, artefact removal, and image enhancement to optimise the input for deep neural networks, which is critical to detect rare anomalies accurately.

By meticulously cleaning the data, researchers can reduce errors and biases that might disproportionately affect the prediction of rare clinical events. This ensures that subsequent steps in the data processing pipeline, such as feature selection and model training, are based on the most accurate and relevant information possible. Effective data cleaning directly contributes to developing reliable predictive models, ultimately aiding in early detection and intervention for rare clinical events.

3.4.2 Feature Selection

After preprocessing, feature selection becomes pivotal in identifying the most informative predictors of rare clinical events. This step reduces dimensionality and enhances model efficiency, which is crucial when dealing with high-dimensional medical data.

Feature selection methods can be broadly categorised into unsupervised and supervised approaches. Unsupervised methods, such as correlation-based feature selection, aim to remove redundant variables by analysing the relationships among independent variables without using the target variable. For instance, correlation coefficients can identify and eliminate highly correlated features, reducing multicollinearity and simplifying the model. Liu *et al.* (2018) utilised correlation-based feature selection to improve the prediction of postoperative complications in colorectal surgery, enhancing model interpretability and performance [60].

While imputation is a common strategy for handling missing data, it is important to note that over-reliance on simplistic imputation methods (e.g., mean imputation) can introduce bias into the data, especially when dealing with rare events.

Supervised methods leverage the target variable to identify and eliminate irrelevant or less important features. These methods are divided into three categories:

- **Wrapper-based methods** use predictive models to assess subsets of features and select the combination that optimises model performance. Recursive Feature Elimination (RFE) is a popular wrapper method. Jiang *et al.* applied RFE with support vector machines to identify critical predictors of complications in patients after mitral valve surgery, resulting in enhanced predictive accuracy [66].

- **Filter-based methods** rely on statistical measures to evaluate the relevance of features with respect to the target variable. Techniques such as the sliding window method, wavelet analysis, and Discrete Wavelet Transform (DWT) have been employed to extract statistical features from time-series data. Mao *et al.* (2019) utilised these methods in machinery fault diagnosis, improving the detection of early faults by capturing essential time-frequency characteristics [67].

- **Intrinsic methods** incorporate feature selection within the model training process. Decision tree-based algorithms, such as Random Forest (RF) and Gradient Boosting (GB), inherently perform feature selection by selecting optimal splits based on feature importance. Huang *et al.* (2024) used decision trees and Lasso regression to identify factors associated with in-hospital mortality in patients with acute aortic dissection, effectively reducing overfitting by penalising less important features [68].

Feature importance methods also rank features according to their influence on model predictions. For example, the Gini importance measure from Random Forests and feature importance scores from XGBoost help identify key factors in rare event prediction. These techniques detect complex relationships and interactions among variables, which are critical in medical data, where interactions may signify important clinical insights.

Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are crucial when dealing with high-dimensional datasets. PCA transforms correlated variables into a set of uncorrelated principal components, capturing the most variance in the data with fewer components. Lever *et al.* (2017) explained how PCA can uncover hidden structures in clinical datasets, aiding in the early detection of rare clinical events and reducing computational complexity [69]. Additionally, t-distributed Stochastic Neighbourhood Embedding (t-SNE) effectively visualises high-dimensional data while preserving local structures, making it useful for analysing complex patterns leading to rare adverse events [70].

By thoughtfully applying these feature selection and engineering techniques, researchers and clinicians can significantly enhance the performance and interpretability of models predicting rare clinical events. This approach paves the way for more accurate and efficient diagnostic and prognostic tools. The choice of specific methods depends on the data's nature, the rare event's complexity, and the computational resources available.

3.4.3 Handling Imbalanced Data

Imbalanced data occurs when the distribution of classes within a dataset is skewed, meaning one class (typically the majority class) significantly outnumbers the other class (minority class). In the context of rare event detection, this imbalance is especially pronounced since rare diseases or events, which constitute the minority class, occur far less frequently than more common conditions. This disparity can cause machine learning models to be biased toward predicting the majority class, resulting in poor performance on the minority class, where the rare events reside.

To address this challenge, various sampling techniques can be used to modify the dataset and achieve a more balanced class distribution, thereby improving the effectiveness of model training and rare event detection.

A common approach is to oversample the minority class. However, duplicating existing instances may lead to overfitting, causing the model to learn patterns specific to the replicated data rather than generalisable features. More sophisticated methods like the Synthetic Minority Over-sampling Technique (SMOTE) generate synthetic examples by interpolating between existing minority class instances. This technique creates a more nuanced augmentation that reduces overfitting risk and helps the model generalise better to unseen data [44]. Adaptive Synthetic Sampling (ADASYN) further refines this by focusing on harder-to-learn examples, tailoring the oversampling process to areas where the model struggles the most.

Further oversampling techniques include Borderline-SMOTE, which generates synthetic samples only in the borderline regions of the minority class, and Cluster-Based Over-Sampling (CBO), which performs oversampling within clusters of the minority class to preserve the inherent structure of the data [71].

Conversely, undersampling reduces the number of majority class instances, potentially discarding valuable information. Techniques like Cluster Centroids aim to preserve essential information by replacing clusters of majority samples with their centroids, thus maintaining the dataset's integrity while addressing imbalance [72].

Advanced undersampling methods further enhance the handling of imbalanced data. Edited Nearest Neighbours (ENN) removes the majority class samples that differ from their nearest neighbours, cleaning the dataset by eliminating noisy or mislabelled instances [73]. Neighbourhood Cleaning Rule (NCL) extends this by removing misclassified majority class samples and certain minority class examples to clean overlapping areas between classes [74]. NearMiss selects the majority class samples close to minority class samples, thereby preserving critical

borderline examples [75].

One-Sided Selection (OSS) combines undersampling with cleaning methods to focus on the most informative majority class examples, reducing the class imbalance while maintaining critical information [76]. Similarity Majority Undersampling Technique (SMUTE) selects majority class samples based on their similarity to minority class instances, aiming to retain those that contribute most to distinguishing the classes [77].

Other sampling strategies include Time Series Subsampling, which involves selecting balanced subsequences from time series data to ensure equal representation of classes over time [78, 79]. Data framing divides audio signals into frames for balanced representation during model training [80].

Uncertainty Sampling is employed in active learning scenarios, where the model selectively queries instances on which it is least sure, often those near the decision boundary. This method focuses labelling efforts on the most informative samples, which is particularly beneficial in rare event prediction [81].

Choice-Based or Endogenous Sampling involves intentionally oversampling the minority class during data collection or analysis. In econometrics, this approach adjusts for the bias introduced by sampling choices and is applicable when the occurrence of rare events can be influenced or observed selectively [82].

Advanced methods such as Importance Sampling (IS) alter the sampling distribution to increase the occurrence of rare events during simulations, enhancing the estimation of rare event probabilities. Standard IS changes the probability measure to focus on critical regions where rare events are more likely to occur. However, this method may require extensive computation, especially in high-dimensional spaces [83]. To address computational challenges and handle complex systems, Mansur *et al.* [84] introduced the Deep Probabilistic Accelerated Evaluation (DeepPrAE) framework, which integrates DL into IS. DeepPrAE efficiently estimates rare event probabilities even in black-box systems, offering statistical guarantees where traditional IS might fail.

Bayesian methods provide another avenue for handling imbalanced data by incorporating prior knowledge with observed data. The Bayesian Updating of Rare Event Probabilities (BUS) framework, discussed by Straub *et al.* [85], updates rare event probabilities using Bayesian inference, combining prior information with new data for improved estimation. Similarly, Rao *et al.* [86] described using ML to solve Bayesian inverse problems, informing the biasing distribution for IS and reducing computational costs in high-dimensional settings.

Other advanced simulation techniques enhance the efficiency of rare event estimation. The Accelerated Failure Identification Sampling (AFIS) method constructs a Gaussian process model to efficiently identify rare failure points, iteratively refining probability estimates with fewer simulations [87]. Additionally, cross-entropy methods optimise the sampling distribution by identifying low-dimensional structures within high-dimensional problems, as Uribe *et al.* (2021) [88] highlights.

Choosing the appropriate sampling technique depends on factors such as the nature of the data, system complexity, and available computational resources. Oversampling methods like SMOTE, ADASYN, and CBO are effective when synthetic data generation can enhance model learning without introducing significant overfitting [44, 89]. Undersampling methods such as ENNs, NCL, NearMiss, OSS, and SMUTE are beneficial when cleaning the majority class and focusing on informative examples is crucial. Importance Sampling and its advanced variants are ideal for probabilistic estimation in complex models, particularly when dealing with high-dimensional data or black-box systems. Bayesian methods are advantageous when prior information or continuous data collection allows for ongoing model updating.

By addressing class imbalance through these sophisticated sampling strategies, models become more sensitive to the minority class, improving detection rates of rare events while maintaining overall performance. Integrating these techniques with effective data preprocessing and feature engineering creates a robust framework for rare event detection in the context of rare diseases. However, it is essential to apply these methods judiciously, considering model interpretability and collaborating with domain experts to ensure clinical relevance and trustworthiness in decision-making processes [90].

3.4.4 Feature Engineering

Feature engineering involves crafting new variables that better capture underlying data patterns, which is particularly important for detecting rare diseases. Leveraging domain knowledge can uncover relationships that are not immediately apparent. For instance, transforming time-series EHRs into meaningful features has significantly improved the prediction of patient outcomes. Harutyunyan *et al.* (2019) developed a benchmark dataset from ICU records and demonstrated that extracting relevant features from time-series data enhanced the prediction of in-hospital mortality and other clinical outcomes [91].

Creating interaction features between environmental exposures and genetic factors improves detecting rare respiratory conditions linked to specific exposures. Patel *et al.* (2010)

highlighted that modelling gene-environment interactions uncovers significant associations in diseases like asthma, where genetics and environment play critical roles [92].

Extracting temporal features from wearable device data has proven effective in detecting rare cardiac events. Analysing irregular heart rate variability over time improves early detection capabilities, showcasing the value of dynamic pattern recognition. Tison *et al.* (2018) showed that wearable devices could passively detect atrial fibrillation by monitoring heart rate patterns, aiding in the early identification of this serious condition [93].

Combining multiple clinical measurements into single risk scores or deriving features from unstructured data, such as text-mining clinical notes, also shows promise in capturing subtle patterns across diverse data sources. Rajkomar *et al.* (2018) discussed how DL models utilising various data types, including unstructured notes, improved predictive accuracy in healthcare settings [94].

3.5 Rare Event Detection Methods

The field of rare event detection has witnessed significant advancements in recent years, driven by the increasing availability of data and the development of sophisticated machine learning (ML) techniques. Data-driven approaches have emerged as powerful tools for identifying and predicting rare events, particularly in healthcare, where early detection of rare diseases can profoundly impact patient outcomes. In this section, we explore various data-driven approaches to rare event detection, focusing on their applications in healthcare and their potential to revolutionise the diagnosis and management of rare diseases.

3.5.1 Statistical Techniques

Statistical techniques provide the foundation for many approaches to rare event detection, focusing on inferring relationships based on assumptions about the data's underlying distribution. These methods typically rely on probability theory and statistical inference to identify patterns indicative of rare events. One of the most widely used statistical techniques in healthcare is logistic regression, which models the probability of a binary outcome (e.g., the presence or absence of a rare disease) as a function of one or more predictor variables. Logistic regression has been successfully applied to identify risk factors for rare diseases, such as genetic and environmental factors associated with rare neurodegenerative disorders [95].

Another important statistical technique is time series analysis, which models and forecasts temporal patterns in data. Time series methods, such as auto-regressive integrated moving average (ARIMA) models and exponential smoothing, have been used to detect rare events in various healthcare applications, including the early detection of infectious disease outbreaks and the identification of adverse drug reactions [96]. These methods are particularly effective in capturing temporal dependencies, which are crucial in understanding how rare events unfold over time.

3.5.2 Machine Learning (ML) Techniques

Machine Learning (ML) techniques differ from traditional statistical methods by not requiring explicit assumptions about data distribution. ML models learn patterns and relationships from data, improving as they are exposed to more examples. These methods can be classified into three main categories: supervised, unsupervised, and semi-supervised learning.

Supervised learning techniques, such as support vector machines (SVM) and decision trees, learn from labelled data, mapping input features to output labels. These models are widely used for rare disease classification, where the goal is to predict the presence or absence of a disease based on patient data. For example, SVM has been used to detect patients with undiagnosed rare diseases based on clinical data [31].

Unsupervised learning techniques, such as clustering and anomaly detection, aim to find hidden patterns in data without labelled examples. These methods are beneficial for detecting rare events when labelled data is scarce or unavailable. Clustering algorithms like k-means and hierarchical clustering can identify rare events by grouping similar instances and spotting anomalies in the data. For example, anomaly detection based on k-means clustering has been used to identify rare heart diseases from clinical data [97].

Semi-supervised learning techniques combine both labelled and unlabelled data to improve model performance. These techniques are instrumental in healthcare applications, where obtaining labelled data for rare events is expensive or time-consuming. For example, co-training and self-training methods have been employed to improve the performance of rare disease classification models by leveraging unlabelled data [98, 99].

3.5.3 Deep Learning (DL) Approaches

DL is a subset of machine learning that focuses on hierarchical feature learning through multilayered neural networks. DL models are highly effective at handling large-scale, complex

datasets, particularly unstructured data such as images, text, and time-series data. DL models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based models have shown remarkable success in rare event detection.

CNNs are highly effective for detecting rare events in medical imaging. These models excel at capturing spatial hierarchies within images, which is crucial for identifying subtle anomalies indicative of rare conditions. For example, CNNs have been successfully applied to detect rare genetic disorders from facial photographs [100].

RNNs, especially Long-Short Term Memory (LSTM) networks, are well-suited for modelling temporal dependencies in sequential data. RNNs have been used in healthcare to detect rare events in time-series data, such as monitoring patient vitals or analysing EHRs. For instance, LSTM networks have been applied to model disease progression in patients with amyotrophic lateral sclerosis (ALS) [101].

Transformer-based models, such as BERT, have revolutionised the field of natural language processing (NLP) and are now increasingly applied to rare event detection in unstructured clinical text data. These models are pre-trained on large corpora and fine-tuned for specific tasks with limited labelled data, making them effective for rare disease detection in clinical narratives and EHR [102, 103].

A key challenge in applying DL to rare event detection is the scarcity of labelled training data. DL models generally require large amounts of data to generalise effectively. Techniques like transfer learning and data augmentation are commonly used to mitigate this challenge.

Transfer learning involves leveraging knowledge from a related task or domain to improve model performance on a target task with limited labelled data. For example, a DL model trained on common diseases can be fine-tuned for rare diseases, allowing the model to leverage learnt features and representations [104].

Data augmentation aims to artificially increase the size of the training dataset by generating modified versions of existing samples. Generative models such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) have been used to generate synthetic examples of rare events, improving the robustness of DL models [105].

In addition, semi-supervised and unsupervised DL approaches, such as autoencoders and deep clustering, have been used to detect rare events with minimal labelled data. These models can learn representations from unlabelled data, proving useful when labelled examples are scarce.

3.5.4 Hybrid Approaches

Hybrid approaches combine multiple data-driven techniques to enhance the performance and robustness of rare event detection models. These approaches leverage the strengths of different methods, addressing challenges such as class imbalance, high dimensionality, and limited labelled data.

A promising hybrid strategy combines DL with transfer learning, which helps improve performance in data-scarce domains. For example, fine-tuning a DL model trained on common diseases with a smaller dataset of rare diseases can enhance detection performance [106].

Another noteworthy hybrid approach integrates ML with expert knowledge to improve model interpretability and reliability. For instance, decision trees can combine data-driven features with expert-defined rules to provide transparent and interpretable models, essential in clinical applications [107].

Hybrid models that combine multiple techniques, including CNN and RNN in healthcare, have shown promising results. These models capture both spatial and temporal patterns in data, leading to improved performance and robustness. For example, a hybrid model using CNNs and RNNs was successfully applied to detect rare adverse drug reactions from EHRs [108].

As the field of rare event detection continues to evolve, hybrid approaches are expected to play an increasingly important role in addressing the unique challenges of rare events. Combining multiple data-driven techniques and incorporating domain-specific knowledge can provide more accurate, reliable, and interpretable models for rare event detection in healthcare and beyond.

3.6 Evaluation Techniques for Rare Event Detection Models

Evaluating the performance of rare event detection models is crucial, given the unique challenges of imbalanced datasets, such as class imbalance and limited positive examples. Specialised evaluation metrics and techniques are necessary to assess the model's effectiveness in accurately identifying rare events.

General evaluation methods, like cross-validation and hold-out validation, should be used cautiously to avoid bias towards the majority class [109]. Rare event-specific evaluation metrics were developed to address these challenges. The Area Under the Receiver Operating Characteristic curve (AUC-ROC) measures the trade-off between true positive and false positive rates, providing a comprehensive view of the model's performance across different classifica-

tion thresholds [110]. Other metrics include precision, which measures the proportion of true positives among all positive predictions; recall, which measures the proportion of true positives among all actual positive instances; the F1 score, which is the harmonic mean of precision and recall; and the Matthews correlation coefficient, which takes into account all four elements of the confusion matrix [111].

Choosing the right evaluation metric depends on the specific requirements of the task. For example, in healthcare applications where missing a rare event (false negative) is costly, a metric that emphasises recall (e.g., F1 score or AUC-ROC) is appropriate. Other considerations when evaluating rare event detection models include stratified sampling to ensure a representative proportion of rare events in the evaluation dataset, using multiple evaluation metrics for a comprehensive understanding of the model's performance, and assessing model calibration using reliability diagrams and calibration plots [112].

Model calibration is critical to evaluating rare event detection models, as it ensures that the predicted probabilities align with the observed event frequencies. Well-calibrated models provide reliable estimates of the likelihood of rare events, essential for decision-making in high-stakes domains like healthcare. Reliability diagrams, also known as calibration plots, are graphical tools used to assess model calibration by comparing the predicted probabilities with the observed event frequencies across different probability bins [113]. These diagrams visually represent the model's calibration performance, with a perfectly calibrated model following the diagonal line. Calibration metrics, such as the Expected Calibration Error (ECE) and the Maximum Calibration Error (MCE), quantify the degree of miscalibration by measuring the average and maximum deviation from perfect calibration, respectively [112].

Techniques for improving model calibration include Platt scaling, which uses logistic regression to transform the model's outputs into well-calibrated probabilities [114], and isotonic regression, which applies a non-parametric monotonic transformation to the predicted probabilities [115]. These methods can help align the predicted probabilities with the observed event frequencies, enhancing the reliability and interpretability of rare event detection models.

Cost-sensitive learning methods are particularly relevant for rare event detection, as they account for the varying costs associated with different types of misclassifications [116]. These methods aim to minimise the total misclassification cost by modifying the training data or the learning algorithm. One approach is to assign different weights to different classes, giving more importance to the rare class. This can be achieved through loss weighting, where the loss function is modified to assign higher weights to the misclassification of rare events. Another

approach is to use focal loss, which down-weights the contribution of easy examples and focuses on complex examples during training. Focal loss effectively handles class imbalance and improves the model's performance on rare events [117].

Fine-tuned evaluation metrics for rare event scenarios include the Precision-Recall curve (PR curve) and Area Under the Precision-Recall curve (AUPR). Precision measures the proportion of true positives out of all positive predictions, while recall (sensitivity) measures the proportion of true positives out of all actual positives. The PR curve visualises the trade-off between precision and recall at different thresholds, and AUPR provides a single value summarising this trade-off. These metrics are more sensitive to class imbalance than AUC-ROC, making them particularly useful for rare event detection tasks [118]. Another metric is the Normalised Discounted Cumulative Gain (NDCG), which assesses the model's ability to rank instances based on their likelihood of being rare events. NDCG is particularly useful when the goal is to prioritise the identification of rare events, as it assigns higher weights to correctly identifying rare instances at the top of the ranked list [119].

3.7 Interpretability and Explainability in Clinical Models

The importance of model interpretability and explainability cannot be overstated in the context of rare event detection, particularly in high-stakes domains like healthcare. Interpretability and explainability are often used interchangeably, but they have distinct meanings. Interpretability refers to the ability to understand how a model works and how it arrives at its predictions. At the same time, explainability provides insights into why a model made a particular decision.

Interpretability techniques in ML are generally categorised into model-intrinsic and post-hoc explanation methods. Model-intrinsic interpretability refers to inherently interpretable models due to their simple structure or the transparency of their decision-making process. These models allow clinicians to directly understand the reasoning behind each prediction from the model's structure. Examples of model-intrinsic interpretability include linear models, decision trees, and rule-based models.

Linear models, such as logistic regression, provide coefficients for each feature, indicating the strength and direction of the relationship between the feature and the outcome. These models allow for easy interpretation and validation against existing medical knowledge in clinical settings. For instance, Lundberg *et al.* (2018) developed a logistic regression model to predict the risk of a rare cardiovascular event that provided clear insights into the most influential risk factors, such as age, blood pressure, and specific biomarkers, enabling clinicians

to understand the model's decision-making process and validate it against established clinical guidelines [120].

Decision trees offer a flowchart-like structure that is easy to follow and understand. Each node in the tree represents a decision based on a feature, leading to a clear path from input to prediction. Clinicians can trace the decision-making process step by step, ensuring transparency. An example of a decision tree model in rare disease diagnosis is the work by Caruana *et al.* (2015), which developed a decision tree to identify patients with a rare genetic disorder based on clinical features and family history. The model's structure allowed clinicians to understand the key decision points and validate them against their domain knowledge [121].

Rule-based models generate a set of if-then rules for prediction, which can be directly interpreted. These models are beneficial for encoding and validating clinical guidelines and protocols. For instance, a rule-based model developed by Tonekaboni *et al.* (2019) for the early detection of a rare infectious disease outbreak provided a set of clear, interpretable rules based on patient symptoms, travel history, and other relevant factors. Public health experts could easily understand and validate these rules, facilitating the model's integration into clinical decision-making processes [122].

Post-hoc explanation methods are used to interpret complex models that are not inherently interpretable. These methods aim to provide insights into the model's behaviour after training. Feature importance methods, such as permutation importance, assess the impact of each feature on the model's predictions. By understanding the most influential features, clinicians can gain insights into the model's decision-making process. In a study by Lipton (2017), permutation importance was used to identify the most critical features in a DL model for detecting rare adverse drug reactions. The analysis revealed that specific patient demographics, medication history, and laboratory results were the most influential factors, providing clinicians with valuable insights into the model's behaviour [123].

Partial Dependence Plots (PDPs) show the relationship between a feature and the predicted outcome, averaged over the distribution of other features. They help visualise how changes in a feature affect the model's predictions. Partial Dependency Plots (PDPs) have been used to interpret ML models in various healthcare applications, including rare disease diagnosis. For example, in a study by Janizek *et al.* (2021), PDPs were used to visualise the relationship between specific biomarkers and the predicted risk of a rare autoimmune disorder, providing clinicians with insights into the model's decision-making process and guiding further clinical investigations [124].

Local Interpretable Model-Agnostic Explanations (LIME) is a popular post-hoc explanation method that explains individual predictions by approximating the complex model locally with an interpretable one. It is particularly useful for describing specific clinical cases, which can aid in understanding unexpected or high-risk predictions. Local Interpretable Model-Agnostic Explanations (LIME) has been applied in various healthcare settings, including rare disease diagnosis. In a study by Ribeiro *et al.* (2016), LIME was used to explain the predictions of a ML model for identifying patients with a rare genetic disorder. The explanations provided by LIME helped clinicians understand the key factors contributing to each patient's diagnosis, facilitating more targeted clinical interventions and improving patient outcomes [125].

SHapley Additive exPlanations (SHAP) is another powerful post hoc explanation method that provides a unified measure of feature importance based on game theory. SHapley Additive exPlanations (SHAP) values offer global and local explanations, indicating how much each feature contributes to each prediction. SHAP has been applied in various healthcare settings, including rare disease diagnosis. For instance, in a study by Lundberg *et al.* (2017), SHAP was used to interpret a ML model for predicting the risk of a rare neurodegenerative disorder. The SHAP values revealed the most influential risk factors, such as age, genetic markers, and environmental exposures, providing clinicians with valuable insights into the model's decision-making process and guiding further clinical research [126].

In the clinical context, the choice of interpretability technique depends on the specific application and the needs of the healthcare professionals. Model-intrinsic methods are preferred when simplicity and direct interpretability are required, such as in developing clinical guidelines or when the model must be transparent for regulatory approval. Post-hoc explanation methods are more suitable for complex models where high predictive accuracy is necessary and the model's intrinsic structure does not lend itself to straightforward interpretation. These methods can help bridge the gap between model complexity and the need for interpretability, allowing clinicians to trust and understand the predictions made by advanced ML models.

However, achieving interpretability and explainability in complex ML models can be challenging. There is often a trade-off between model performance and interpretability, as simpler, more interpretable models may not capture the intricate patterns and relationships in the data. Researchers and practitioners must strike a balance between model complexity and interpretability, depending on the application's specific requirements. For instance, in rare disease diagnosis, where the consequences of misdiagnosis can be severe, it may be more important to prioritise model interpretability over marginal gains in predictive performance. On the other

hand, in applications where the cost of false negatives is high, such as in the early detection of rare cancer subtypes, more complex models with higher predictive accuracy may be preferred, with post hoc explanation methods used to provide insights into the model's decision-making process.

Striking the right balance between model complexity and interpretability is critical in developing clinical decision support systems for rare disease diagnosis. It requires close collaboration between ML experts, clinicians, and other stakeholders to ensure that the models are accurate, transparent, trustworthy, and aligned with clinical knowledge and ethical principles.

3.8 Translating Models into Clinical Practice

Translating interpretable and explainable AI models into clinical practice is a complex, multidisciplinary endeavour that requires close collaboration among data scientists, healthcare professionals, IT experts, and policymakers. This process involves not only the technical aspects of model development and deployment but also the integration of these models into existing clinical workflows, acceptance by healthcare professionals, and adherence to regulatory standards.

Developing AI models for clinical decision support is a challenging task that requires a deep understanding of the clinical domain, the limitations of the available data, and the ethical implications of AI-driven decisions. To ensure the successful translation of these models into practice, adopting a human-centred design approach that involves healthcare professionals and patients throughout the development process is essential. Key principles of human-centred design that should be considered include empathy, collaboration, iteration, transparency, and adaptability.

Several interpretable and explainable AI models, such as the Early Warning Score (EWS) system and the IDx-DR diabetic retinopathy diagnostic system (IDx-DR) system, have been successfully integrated into clinical practice, although with some limitations. The Early Warning Score (EWS) system uses ML algorithms to predict patient deterioration based on vital signs and other clinical data, offering explanations for its predictions. However, clinicians may find these explanations difficult to interpret, and there is a risk of over-relying on the system's predictions without accounting for other clinical factors. Similarly, the IDx-DR system, an AI-based tool for detecting diabetic retinopathy from retinal images, provides interpretations of its findings [127]. While it has received regulatory approval, its explanations may not fully cap-

ture the underlying disease mechanisms, and there is a potential for false positives or negatives in certain cases.

These examples highlight the challenges of translating interpretable and explainable AI models into clinical practice. Key issues to address include the complexity of explanations, over-reliance on predictions, integration challenges, incomplete understanding of disease mechanisms, and the potential for false positives and negatives.

Moreover, general challenges and gaps in the translation process include ethical and legal issues surrounding accountability, the lack of contextual information available to AI models, and the risk of over-reliance on diagnoses made by AI. These challenges are further amplified when dealing with rare clinical events or diseases, as the scarcity of data and limited expert knowledge make it harder to develop and validate AI models that can accurately predict and diagnose these events. The heterogeneity of rare diseases and the variability in their presentation and progression also pose significant challenges to the generalisability and adaptability of AI models.

Addressing these challenges requires ongoing research, multidisciplinary collaboration, and a focus on human-centred design principles to ensure these tools are effective and trustworthy in real-world clinical settings. Ensuring that model explanations are clinically meaningful, actionable, and easily understood by healthcare professionals is a critical goal, but harder to achieve when dealing with less common conditions. Improving the generalisability and adaptability of these models across different healthcare settings and patient populations is an ongoing area of research, with rare diseases presenting unique challenges due to their inherent variability and the limited availability of data and expertise.

In healthcare systems, the balance between false positives and early detection is crucial, especially for rare diseases. While minimising false positives is essential to avoid unnecessary interventions and patient anxiety, the ability to catch rare diseases early can significantly improve patient outcomes and reduce healthcare costs in the long run. However, achieving this balance is particularly challenging for rare conditions, as the limited data and understanding of these diseases make it harder to develop AI models that can accurately distinguish between true positives and false positives.

Ultimately, successfully translating interpretable and explainable AI models into clinical practice requires a nuanced understanding of the trade-offs between false positives and early detection, particularly in rare diseases. By engaging healthcare professionals, patients, and other stakeholders in the development process, AI researchers can work towards finding the

optimal balance that maximises patient benefits while minimising unintended consequences. This collaborative approach, combined with ongoing research and investment in interpretability and generalisability, is crucial for realising the full potential of AI in healthcare and improving patient outcomes, especially for those affected by rare conditions.

3.9 Research Gaps and Future Directions

Despite the significant progress in rare event detection, several research gaps and challenges remain. Addressing these gaps is crucial for advancing the field and developing more effective and reliable rare event detection models.

One of the main gaps in the literature is the lack of standardised datasets for rare event detection, which makes it difficult to compare the performance of different models and techniques effectively. Future research should focus on developing comprehensive, domain-specific datasets that accurately represent the challenges of rare event detection across various fields, including healthcare, finance, and cybersecurity. The development of such datasets will be a key focus in the upcoming chapters, particularly in Chapter 7, where we explore the creation of a rare disease dataset tailored explicitly for the UK healthcare system.

Emerging research trends focus on leveraging advanced learning methods for rare event detection, such as DL. These methods have shown promising results in various domains, but their application in rare event detection is still in its early stages. Future research should investigate DL architectures optimised for imbalanced datasets and transfer learning techniques to leverage knowledge from data-rich domains to improve rare event detection in data-scarce areas. This thesis will explore the use of DL techniques, namely self-attention mechanisms (Chapter 5) and graph neural networks (Chapter 6), in moderate and severely imbalanced datasets.

Another promising direction is the integration of domain knowledge into prediction models, which can help improve both the interpretability of the model and the accuracy of predictions. Future research should focus on developing frameworks for seamlessly incorporating expert knowledge into ML models, creating hybrid models that combine data-driven approaches with theory-based models. Chapter 6 proposes a novel approach for infusing medical ontologies into language models to improve their understanding of infrequently coded conditions and their relationships.

As data generation continues to accelerate, developing methods for streaming data becomes increasingly important. Future research should address algorithms capable of adapting to changes in the underlying distribution of rare events over time. While this is not a primary

focus of this thesis, we will briefly discuss the potential of online learning algorithms for rare disease detection in the context of the UK healthcare system in Chapter 8.

Given the scarcity of labelled data in many rare event scenarios, unsupervised and semi-supervised learning methods offer promising avenues for future research. This includes developing robust anomaly detection algorithms to identify novel types of rare events. Chapter 7 will delve into applying unsupervised learning techniques for distinguishing rare cardiac diseases with similar presentations.

Future research in rare event prediction is directed towards developing more sophisticated models, leveraging advancements in computational power, and expanding data availability. Researchers are also focusing on improving the interpretability and usability of models to facilitate their adoption in real-world settings. Throughout this thesis, we will emphasise the importance of model interpretability and usability, particularly in the context of clinical decision support systems for rare disease diagnosis (Chapters 7 and 8).

It is important to continue researching and developing new techniques for rare event detection. By addressing these identified gaps and challenges, we can advance the field and contribute to more effective, reliable, and ethically sound rare event detection models that can be confidently applied across various domains.

3.10 Summary

This chapter provided a comprehensive overview of data-driven approaches in rare event detection, focusing on healthcare applications. It began with examining the unique characteristics and challenges of rare event data, including class imbalance, high dimensionality, temporal complexities, and the importance of contextual information. These challenges necessitate specialised data preprocessing and feature engineering techniques to improve data quality and model performance.

Next, the sections explored various data-driven methodologies for rare event detection, highlighting the role of statistical techniques, ML, and DL approaches. It was discussed how ML techniques, including supervised, semi-supervised, and unsupervised learning, have been applied to detect rare diseases and adverse events in healthcare. The advancements in DL, particularly convolutional neural networks and recurrent neural networks, have shown promise in capturing complex patterns in high-dimensional data. There was an emphasis on the potential of transfer learning to leverage knowledge from data-rich domains and improve rare event detection in data-scarce areas.

The evaluation of rare event detection models was discussed, focusing on specialised metrics and techniques to assess model performance accurately. The importance of model calibration was highlighted, and tools that ensure predicted probabilities align with observed event frequencies were introduced. These evaluation techniques are crucial for developing reliable models that can be confidently applied in high-stakes domains like healthcare.

A significant emphasis was placed on interpretability and explainability in clinical models. We discussed the importance of model-intrinsic interpretability and post-hoc explanation methods, providing examples specific to the healthcare domain. The trade-off between model performance and interpretability was explored, underscoring the need to strike a balance depending on the application's requirements. Practical and trustworthy clinical decision support systems for rare disease diagnosis can be further developed by leveraging the strengths of interpretable models and advanced explanation techniques.

Furthermore, we delved into the challenges of translating these models into clinical practice. Adopting a human-centred design approach was highlighted as essential for successful integration, involving healthcare professionals and patients throughout development. We discussed practical aspects such as understanding user needs, fostering stakeholder collaboration, iteratively refining models based on feedback, ensuring transparency in AI decision-making, and adapting models to changing clinical contexts. The EWS and IDx-DR systems demonstrated the potential of implementing interpretable and explainable AI models in clinical settings.

Despite progress, several research gaps and challenges remain in rare event detection. We identified the need for standardised datasets, improved uncertainty quantification, advanced learning methods tailored for imbalanced data, and the integration of domain knowledge into prediction models. Addressing these gaps is crucial for advancing the field and developing more effective, reliable, and ethically sound rare event detection models.

Throughout our work, these challenges in the context of the UK healthcare system will be explored with a focus on rare disease diagnosis. Data scarcity (Chapter 4) will be discussed, as will model interpretability (Chapters 7 and 8) and the integration of domain knowledge (Chapter 6), leveraging advanced ML techniques such as DL (Chapters 5 and 6), Bayesian methods (Chapter 4) and unsupervised learning (Chapter 7). By developing novel approaches and frameworks, our work contributes to advancing rare event detection and improving patient outcomes.

Moving forward, it is essential to prioritise the development of robust, reliable, and eth-

ically sound rare event detection models that can be confidently applied in clinical practice. By addressing the key challenges of interpretability, model deployment, and stakeholder engagement, it is possible to harness the power of AI to predict better and manage rare events, ultimately improving outcomes for individuals and society. This chapter serves as a foundation for the remainder of this thesis, highlighting the key challenges and opportunities in rare event detection and setting the stage for detailed investigations to follow.

Chapter 4

Navigating Imbalance and Missing Data in Clinical Settings

Contents

4.1 The silent killer

Sepsis is a life-threatening condition that arises when the body's response to infection causes widespread inflammation and organ dysfunction [128]. It is a silent killer, often progressing rapidly and leading to high mortality rates if not promptly and effectively treated [129]. In 2017, an estimated 48.9 million sepsis cases were reported worldwide, and it was responsible for 11 million deaths [129]. Sepsis is a concern, especially in ICUs.

A study examining 30 years of sepsis trends in the UK reported an increase in the proportion of sepsis cases in the ICU from 7.6% between 1988 and 1990 to 19.6% between 2017 and 2019, an increase of 2.6-fold [130].

Patients in the ICU are especially vulnerable to sepsis due to their underlying conditions, invasive procedures, and exposure to multiple risk factors. The rapid progression of sepsis in these patients can lead to severe complications, such as septic shock, which is associated with a mortality rate of up to 50% [131,132]. The complexity of sepsis, combined with its heterogeneous presentation and the imbalanced nature of clinical data, presents unique challenges for developing predictive models that are both accurate and clinically applicable.

Early detection of sepsis is crucial for improving patient outcomes and reducing the burden on healthcare systems [133]. However, the imbalanced nature of clinical data, where non-septic instances far outnumber septic cases, poses challenges for developing accurate predictive models. Addressing this issue requires sophisticated ML techniques and a deep understanding of the clinical context and dynamics of ICU data.

This chapter further explores ML techniques for early detection of sepsis within the ICU, with the use of the Medical Information Mart for Intensive Care-IV (MIMIC-IV) dataset. The primary focus of this study is to enhance the interpretability and reliability of ML models designed for early sepsis prediction during an ICU stay. By developing accurate and clinically applicable models, we contribute to the knowledge base, provide insights that could improve sepsis outcomes and present a diagnostic support tool to facilitate early sepsis detection.

4.2 Diagnostic Scores and Pathways and their pitfalls

Diagnosing sepsis can be challenging due to its complex and heterogeneous presentation [130]. Various diagnostic scores and pathways have been developed to aid in the identification of sepsis. However, a systematic review evaluating the performance of the individual scores found that it ranged between poor and acceptable at best [134]. Different combinations of these scores may be better for accurately detecting sepsis.

Standard sepsis detection scoring systems, such as the Sequential Organ Failure Assessment (SOFA) and the Systemic Inflammatory Response Syndrome (SIRS) criteria, are easily accessible and calculated. Still, each has limitations [135]. The SOFA score relies on organ dysfunction and measurements from the ICU, which is less valuable when detecting sepsis in its earlier stages [135]. SIRS, on the other hand, was critiqued for being non-specific.

Moreover, these diagnostic scores and pathways do not capture the full spectrum of sepsis presentations. They do not account for atypical signs or symptoms, especially among the elderly population, or the dynamic nature of the condition [136]. Hence, relying solely on these tools can lead to missed or delayed diagnoses, compromising patient care and outcomes [137].

EHRs allows for more efficient patient monitoring and early identification of sepsis by leveraging key biomarkers and clinical data. ML algorithms can be applied to EHRs data, enabling automated and real-time sepsis prediction without additional clinician input.

A key challenge when using ML for any analysis is the imbalanced nature of clinical datasets, where non-septic instances far outnumber septic cases, as noted in these studies [138–140]. This can lead to biased models that fail to accurately identify critical sepsis cases. Addressing this requires sophisticated techniques to balance the data and adjust the algorithm to ensure the models' accuracy and sensitivity in detecting early signs of sepsis in the imbalanced dataset.

Another challenge is the scepticism of healthcare professionals to trust decisions made by ML tools, which are usually unfamiliar with the underlying algorithms and how such tools achieve certain conclusions. These can be addressed by incorporating interpretability and explainability into the design of ML tools that aid in decision-making [141].

In sum, diagnostic scores and pathways previously aided in sepsis detection have unique limitations. The complexity and heterogeneity of sepsis presentation and the imbalanced nature of data within EHRs require a more comprehensive approach that leverages the power of ML. Therefore, we aim to develop clinically aligned models that consider a wide range of data and provide interpretable results to improve early sepsis detection and support clinical decision-

making in the ICU as a decision support tool.

4.3 Confronting data gaps in critical care

Recent statistics show that between 25%-30% of individuals with sepsis die from it and hospital mortality from septic shock is between 40% -60% [128]. Despite improvements in diagnostic criteria over time, diagnosing sepsis promptly, especially in its early stages, remains a challenge [130]. This is primarily due to the lack of formal diagnostic criteria, especially for early sepsis, as it is often multifactorial and heterogeneous in presentation [142]. This resulted in most sepsis research focusing on validating existing diagnostic criteria and discovering diagnostic markers [137, 143–148].

Misra et al. identified lactic acid as a key diagnostic marker where increased concentrations were associated with progression to septic shock and death [149, 150]. Misra, Wardi and researchers [151, 152] achieved high accuracy performances of at least 85% in their complex algorithms.

Besides identifying key diagnostic features, it is vital to provide a way for healthcare professionals who are unfamiliar with the models to understand how conclusions were reached [141]. This concern has been addressed briefly in recent studies using sophisticated libraries, including SHAP. SHAP is a quantitative measure of feature importance by comparing accuracies in prediction with and without the feature.

Li et al. pioneered explainability tools by using SHAP values to display the importance of selected key features and provide visual quantitative insights on their effect in the model's decision for the patient's diagnostic status. An example showcases the combined effect of temperature and ICU length of stay (ICU LOS) on the risk of diagnosing sepsis. It suggests that the risk of sepsis is relatively low at lower temperatures, regardless of ICU Lenght Of Stay (LOS). The opposite is true, with longer ICU LOS indicating higher susceptibility to sepsis [147]. The insight demonstrated the usefulness of including visual aids and SHAP to help algorithm users better understand the models' decisions.

The examples above illustrate that many studies have contributed valuable insights into sepsis diagnosis and intuitive decision-making tools for early sepsis detection [150, 153, 154].

However, investigation into the data preparation methods shows that most studies apply the same set of methods to handle missing data across the complete set of features; some removed records which had a proportion of missing values above a predefined threshold, in which Misra et al. excluded variables with 40% or more missing values. Yuan et al. ex-

cluded records with missing vital sign values, and Bloch et al. and Saqib et al. only used fully completed records [150, 155–157]. Few did not apply any imputation, while others used forward-fill variations or imputation to complete the dataset [158–160]. While these methods apply to well-balanced datasets, they may not be the most suitable for rare diseases, as they incorrectly treat all measures equally. For instance, vital signs like heart and respiratory rate, blood pressure and body temperature are taken more frequently than laboratory tests, which are ordered circumstantially.

Therefore, this study aims to answer the following questions:

1. How does a clinically informed model perform compared to existing models in sepsis detection?

2. How clinically aligned are the most impactful features generated by the model?

4.4 Designing a clinically aligned approach to modelling ICU data?

4.4.1 Data source

This study uses data from the Medical Information Mart for Intensive Care IV (MIMIC-IV version 1.0) database from 2008 to 2019 [161]. MIMIC includes over 40,000 ICU patients admitted to Beth Israel Deaconess Medical Centre (BIDMC) and has a 1:3 ratio of sepsis to non-sepsis patients. We accessed MIMIC through Google BigQuery using data extraction codes from `https://github.com/MIT-LCP/mimic-iv`. Modifications were made to include additional parameters in the blood gases, cardiac markers and enzyme assay tables. The study used 137 features in the dataset, including administrative details, vital signs, comorbidities and other diagnostic tests.

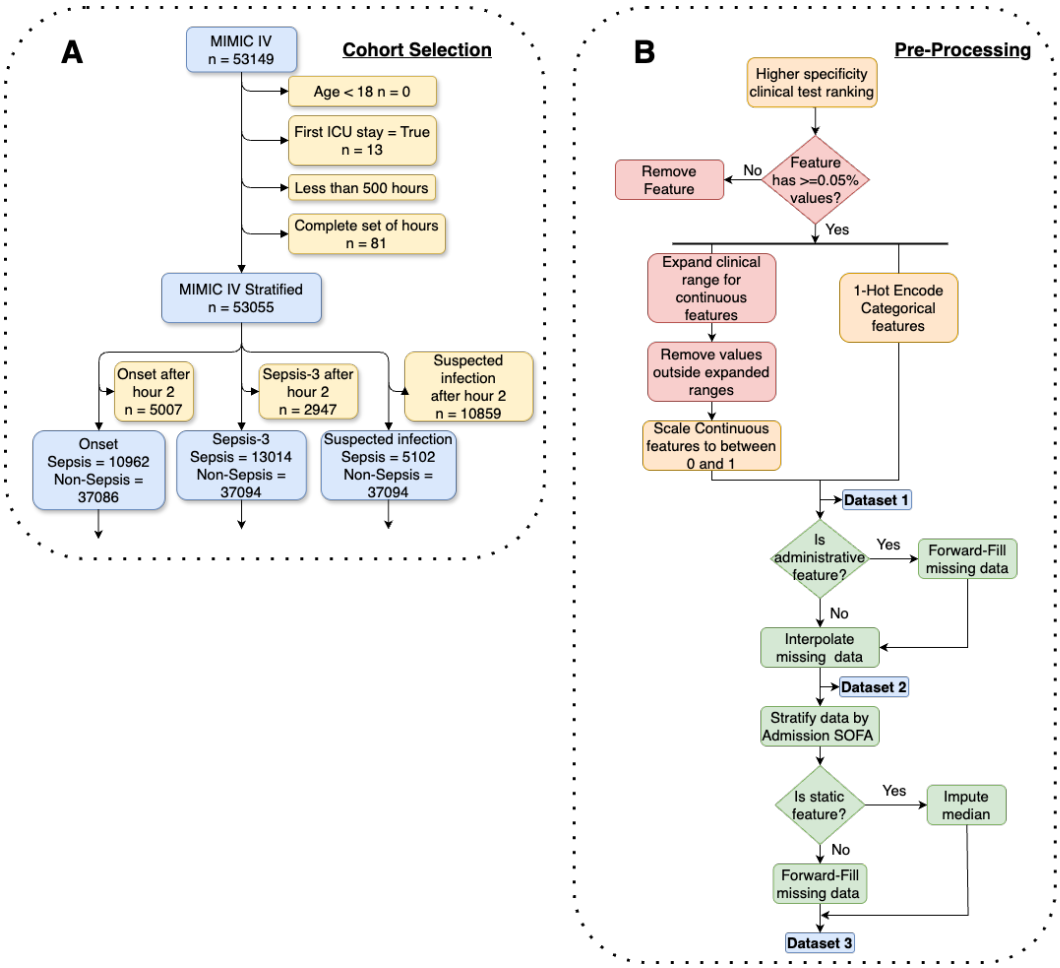


Figure 4.1: Flow chart illustrating the data preparation process. (A) Cohort selection where yellow boxes represent exclusion criteria and blue boxes indicate successive versions of the complete dataset until stratified by diagnosis point. (B) Preprocessing, where orange boxes refer to data modification steps, red boxes indicate data removal, and green boxes indicate an imputation step. Throughout each imputation stage, a version of the dataset (blue boxes) is extracted later to assess the impact on performance and clinical validity.

4.4.2 Case Ascertainment

Sepsis-3 criteria were used for case definition and to understand prodromal symptoms [162]. Records were included if they met the following criteria:

- The patient is aged 18 and above:

  By restricting the cohort to adults, the study focuses on the population most likely to encounter sepsis in ICU settings. Paediatric sepsis differs in presentation, management,

and prognosis, requiring distinct protocols for assessment [163]. Limiting the study to adults ensures the findings apply to most ICU patients [162].

- The episode was an index ICU stay:
  Including only the first ICU admission per patient minimises confounding from recurrent admissions, which can involve multiple comorbidities, complications, and treatments that may obscure the early markers of sepsis. This approach helps maintain the homogeneity of the study cohort and reduces bias due to the complexity of multi-episode ICU admissions [162].

- Sepsis developed after 2 hours of ICU admission:
  Sepsis typically develops after admission to the ICU, often as a result of invasive procedures, infections, or underlying conditions worsened by critical care treatments. Excluding patients who already had suspected sepsis upon admission isolates those in whom sepsis developed as a consequence of ICU or previous care. This avoids confounding factors related to patients suspected of having sepsis at the time of admission [146, 162].

- The episode was 21 days or less (i.e., 500 hours between ICU admission and hospital discharge):
  Sepsis is most commonly acute, and limiting the study to patients with ICU stays $\leq$21 days ensures a focus on early-stage sepsis. Longer ICU stays often involve multi-organ failure and chronic conditions, making sepsis diagnosis more complex. By capping the stay at 21 days, we capture patients with a typical sepsis trajectory, providing a more accurate reflection of early sepsis detection. This 21-day threshold is consistent with ICU sepsis definitions and helps to avoid the complexity of long-term ICU admissions, where sepsis may be secondary to other complications [146, 162].

4.4.3 Data pre-processing

Diagnosis points of sepsis were chosen using the Third International Consensus Definitions for Sepsis (sepsis-3), also used by Nemati *et al.* [146].

The diagnostic times identified are the following (Figure 1):

- Suspected infection ($t_{suspicion}$): Earliest time an antibiotic was given, followed by a blood culture being taken no more than 24 hours after a blood culture was taken, then an an-

tibiotic being administered within 72 hours. This diagnostic point reflects the clinician's suspicion that infection is present and needs to be treated.

- Sepsis-3 (t_{sepsis}): an episode of suspected infection with a two- or more-point change in the SOFA score (t_{SOFA}) from up to 24 hours before to up to 12 hours after the $t_{suspicion}$ (t_{SOFA}+24h>$t_{suspicion}$>t_{SOFA}-12h).

  This refers to the most recent sepsis criteria that require organ dysregulation in addition to a suspected infection for diagnosis.

- Onset (t_{onset}): the earliest time point, judged in retrospect, once sepsis is diagnosed, that organ dysfunction was assumed to be caused by sepsis; either the point at which infection was suspected ($t_{suspicion}$) or the point at which t_{SOFA} changed if that is earlier.

Our work focuses on detecting t_{sepsis} as it resembles the current gold standard. Predictive performances of the algorithm on $t_{suspicion}$ and t_{onset} were given for completeness and to facilitate comparison with existing literature.

4.4.4 Data completeness

Clinical Marker	Category	Rank 1	Rank 2	Rank 3
Bicarbonate	Laboratory	Blood Gases	Renal Function	
Chloride	Laboratory	Blood Gases	Renal Function	
Respiration Rate	Clinical Observations/Ventilator Settings	Ventilator	Vital Signs	
Sodium	Laboratory	Blood Gasses	Renal Function	
Temperature	Clinical Observations	Blood Gasses	Vital Signs	
Oxygen Saturation	Laboratory/Clinical Observations	Vital Signs	Blood Gases	
Haematocrit	Laboratory	Complete Blood Count	Blood Gases	
Haemoglobin	Laboratory	Complete Blood Count	Blood Gases	
Fraction of Inspired Oxygen (FiO2)	Laboratory/Ventilator Settings	Ventilator	Blood Gases	
Glucose	Laboratory/Point of care	Plasma glucose (lab)	Blood Gases	Point of care glucometer
PEEP	Clinical Settings/Laboratory	Blood Gases	Ventilator	
Potassium	Laboratory	Blood Gases	Renal Function	
Oxygen flow	Clinical Observations	Ventilator	Blood Gases	
Tidal Volume	Ventilator Settings	Blood Gases	Ventilator	

Table 4.1: Feature ranking of biomarker measurements and feature type. Rank 1 being the most clinically specific and Rank 3 being the least.

Having defined diagnosis points, the focus was on improving data quality. Initially, biomarkers with more than one measurement technique were ranked by clinical specificity 4.1. For

example, blood glucose measured with a glucometer is taken more frequently. Still, it is less accurate than using an arterial blood gas (ABG) analyser [164] and much less precise than laboratory-tested samples.

Where at least two measures are available, laboratory results would supersede an ABG result, which would supersede the glucometer measurements.

Missing data, on the other hand, requires more interventional methods. The most popular methods of handling missing data are combinations of 1) interpolation with mean imputation [144, 158] and 2) forward-fill or carry-forward imputation with mean imputation [147, 154].

Extending from these methods, features with less than 0.05% available values were excluded (Figure 4.3). Additionally, most clinical tests were only requested at set intervals or upon clinical suspicion [165], a frequency analysis was done for insights on data missingness. Continuous variables were then filtered to retain between predefined clinically possible ranges. Administrative features were forward-filled while others were interpolated. Refer to Appendix table A.1

Three test datasets were created using varying methods of handling missing data:

1. Numerical values were scaled to 0-1, and categorical variables were converted into individual columns with binary vectors (one-hot encoded).

2. Preceding method: If there were two or more values for a biomarker in a patient'ICU stay, the missing values in the middle would be a product of the interpolation between both existing values.

3. The preceding method was used with the addition of stratifying patients by their admission SOFA score. Excluding administrative and categorical variables, features were assigned into static SD $\leq$ 0.1 or dynamic categories. 0.1<SD $\leq$ 1. Static features were imputed with the admission SOFA group median scores, while dynamic features were forward-filled.

Subsequently, three diagnosis points were allocated to each dataset: 1) At the original diagnosis time point according to Sepsis 3 criteria, 2) 6 hours before and 3) 12 hours before the original diagnosis time. This ensures that data patterns identified were not from the natural variation in missingness between septic and non-septic patients, also allowing for early sepsis diagnosis of up to 12 hours. The non-imputed full-feature data with the original diagnosis point was used as the baseline dataset.

4.4.5 Machine Learning Model Selection and Evaluation

Our work used ML techniques to predict healthcare outcomes, leveraging increasingly available complex clinical datasets. ML methods offer substantial advantages in predictive performance and identifying previously undetected subpopulations, which is especially beneficial for improving patient care and outcomes in clinical environments [166]. Below is an outline of our model development process, including the rationale for the choices and relevant literature that informed these decisions.

Hyperparameter Tuning

The first step in model optimisation involved hyperparameter tuning, which was carried out using `Optuna` with `Hyperband` optimisation. Hyperparameter tuning is crucial for optimising model performance, as emphasised in multiple studies, including Stevens et al. [167], who highlight the need for careful hyperparameter selection in clinical ML applications to ensure generalisability. `Optuna` was selected for its ability to efficiently explore large hyperparameter search spaces using Bayesian optimisation, a method known for balancing exploration and exploitation during the search process [168]. Additionally, `Hyperband` dynamically allocates resources during the optimisation, making it particularly efficient in selecting hyperparameters when computational resources are limited [169].

For the classifier, `XGBoost` was chosen for its strong performance with structured data, particularly in clinical research. `XGBoost` has been shown to outperform other models in numerous healthcare prediction tasks, as discussed by Venkatesan and Yamuna [170]. The specific hyperparameters optimised were learning rate, maximum tree depth, and the number of estimators, which are crucial for achieving optimal model performance in clinical applications. These hyperparameters were optimised using logarithmic loss as the objective function to ensure the best possible model performance for classification tasks in clinical applications.

The hyperparameter tuning process was conducted over 100 iterations to find the optimal set of hyperparameters. During this process, stratified K-fold cross-validation with 5 folds was employed to ensure that the data was evenly distributed across each fold while maintaining the class distribution in each subset. This approach provided a robust evaluation of the model's performance during the hyperparameter optimisation, helping to prevent overfitting by validating across multiple data splits. The results from the cross-validation guided the selection of the best hyperparameter configuration for further model training and evaluation.

Model Training and Validation

After hyperparameter tuning, the final model was trained using a train-validation-test split. The dataset was divided into 60% for training, 20% for validation, and the remaining 20% for testing. This split allowed the model to learn from a sufficient portion of the data while retaining separate subsets for validation and testing. The validation set was used to tune the model during training and assess its performance, ensuring it generalised well to unseen data and reducing the risk of overfitting during training.

Test Set Evaluation

The trained model was then evaluated on the remaining 20% of the dataset, serving as the test set. The test set provided an unbiased estimate of the model's ability to generalise to new, unseen data, which is essential in clinical applications, where predictions must be made for future patient populations [167]. Performance metrics such as accuracy, precision, recall, and F1-score were computed to assess the model's effectiveness, particularly given the class imbalance often present in healthcare datasets [171]. These metrics were chosen for their ability to evaluate different aspects of model performance, especially in imbalanced datasets, where traditional accuracy measures may not reflect true predictive power.

Model Interpretability

Given the importance of model interpretability in clinical decision-making, `SHAP` (Shapley Additive Explanations) values were employed to elucidate how each feature influenced the model's predictions [172]. `SHAP` values, based on cooperative game theory, provide a clear explanation of the contribution of each feature to the model's output [126]. Stevens et al. [167] discuss the importance of interpretability for clinical practitioners, as ML models must be transparent for integration into clinical decision-making. This transparency allows healthcare providers to trust and understand the model's predictions, which is critical for adoption in clinical practice.

Additional Considerations and Rationale

The choice of `XGBoost` was driven by its predictive performance and strong interpretability features, making it a practical choice for healthcare applications where decision support systems must be transparent. Using `SHAP` aligns with the guidelines presented by Stevens et

al. [167], who emphasise the importance of making model outputs understandable to clinical end-users. The validation of these results using external test sets was also a key strategy to ensure that the model's performance was not overestimated due to overfitting, per best practices for ML evaluation [173].

Feature Selection and Interpretation

The model was trained first on the complete feature set, and performance metrics were recorded from predictions on the test set. Features with an importance gain of 0.1 and above were extracted using the XGBoost importance function. Next, features were removed sequentially in descending order of importance gain, and a new model was trained. This was repeated until a sharp decrease in accuracy was observed. The corresponding performance metrics were then recorded to showcase the diagnostic importance of a reduced set of features in comparison with other studies that utilised a subset of frequently recorded features [150, 153, 154, 158, 174].

A common approach to select key features is through the importance gain ranking, which determines the impact of each feature by its effect on correctly classified samples. However, the usefulness of importance gain is being debated. A strong criticism was its inconsistency when looking for general trends in the data. This means additional tools or alternate metrics should be considered to evaluate feature importance. Tree SHAP, part of the Shapley library (SHAP) [120], is mathematically equivalent to averaging the differences in predictions over all possible orderings of the features, rather than just the ordering specified by their position in the tree, which is the approach of importance gain. This research aims to validate the clinical relevance of its findings with the use of Tree SHAP as done in Liu et al. [147].

4.5 Highlighting known patterns and novel markers

		Missing	Overall	Control	Case
Number of Patients			48048	37086	10962
Age, mean (SD)		0	63.7 (17.6)	63.3 (17.7)	65.1 (17.1)
Ethnicity, n (%)	**AMERICAN INDIAN/ALASKA NATIVE**	0	91 (0.2)	68 (0.2)	23 (0.2)
	ASIAN		1431 (3.0)	1112 (3.0)	319 (2.9)
	BLACK/AFRICAN AMERICAN		5454 (11.4)	4273 (11.5)	1181 (10.8)
	HISPANIC/LATINO		1906 (4.0)	1471 (4.0)	435 (4.0)
	OTHER		2186 (4.5)	1683 (4.5)	503 (4.6)
	UNABLE TO OBTAIN		482 (1.0)	359 (1.0)	123 (1.1)
	UNKNOWN		3965 (8.3)	2995 (8.1)	970 (8.8)
	WHITE		32533 (67.7)	25125 (67.7)	7408 (67.6)
Gender, n (%)	**F**	0	21938 (45.7)	16982 (45.8)	4956 (45.2)
	M		26110 (54.3)	20104 (54.2)	6006 (54.8)
Weight, mean (SD)		2153	82.4 (25.9)	82.0 (25.6)	83.4 (27.1)
First hospital stay, n (%)	**False**	0	10849 (22.6)	7949 (21.4)	2900 (26.5)
	True		37199 (77.4)	29137 (78.6)	8062 (73.5)
Max heart rate, mean (SD)		11	103.1 (18.2)	100.8 (17.6)	110.7 (17.8)
Charlson comorbidity index, mean (SD)		0	5.3 (3.0)	5.1 (3.0)	5.8 (3.0)
SOFA 24h, mean (SD)		0	3.6 (2.9)	3.1 (2.6)	5.3 (3.2)

Table 4.2: MIMIC-IV population descriptive statistics, stratified by septic and non-septic patients.

4.5.1 Population analysis

The cohort comprised 48048 patients, 26.0% of whom had a sepsis diagnosis according to sepsis-3 criteria during their first ICU stay. Table 4.2 provides key information on the study population used to train the model and validate findings. The ratio of sepsis to non-sepsis instances at each predetermined time point was similar. There were 41.4% septic instances at the sepsis-3 diagnostic point, which rose to 43.2% and 44.0% at 6 and 12 hours before.

In terms of ethnicity, the majority of the cohort were White (67.7%) followed by Black-/African American (11.4%), with the smallest minority as American Indian/Alaska Natives (0.2%). The ethnicity distribution between cases and controls was similar.

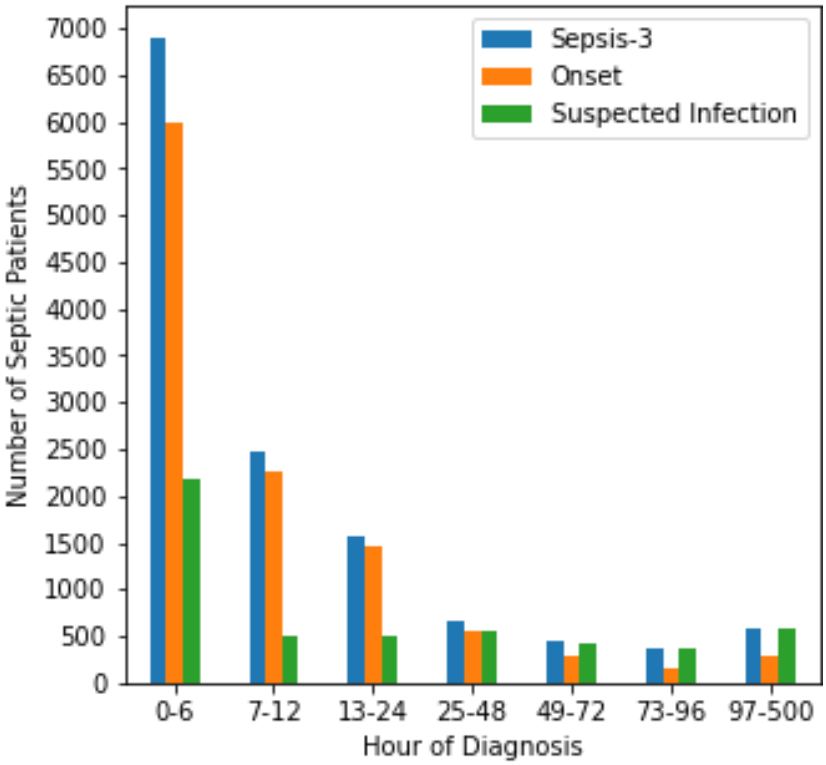


Figure 4.2: Count of septic patients by hour intervals and sepsis diagnostic criteria.

Of note, the average age of 63.7 years suggests that most of the population in the study cohort were elderly patients. Septic patients were on average 2 years older than non-septic patients. Cases had higher maximum values for heart rate, systolic and diastolic blood pressure, comorbidity indexes and SOFA scores compared to controls. Patients with multiple hospital admissions before their first ICU triage were more likely to suffer from sepsis than those who were triaged to ICU during their first hospital admission. Figure 4.2 shows that a notable 82.1% of cases developed sepsis within 24 hours of their ICU admission. Half of the sepsis diagnoses occurred within the first 6 hours of ICU stay, and the second half in the subsequent 6 hours.

The following section discusses the distribution of missingness found in the data and during the data processing stage.

4.5.2 Missing data

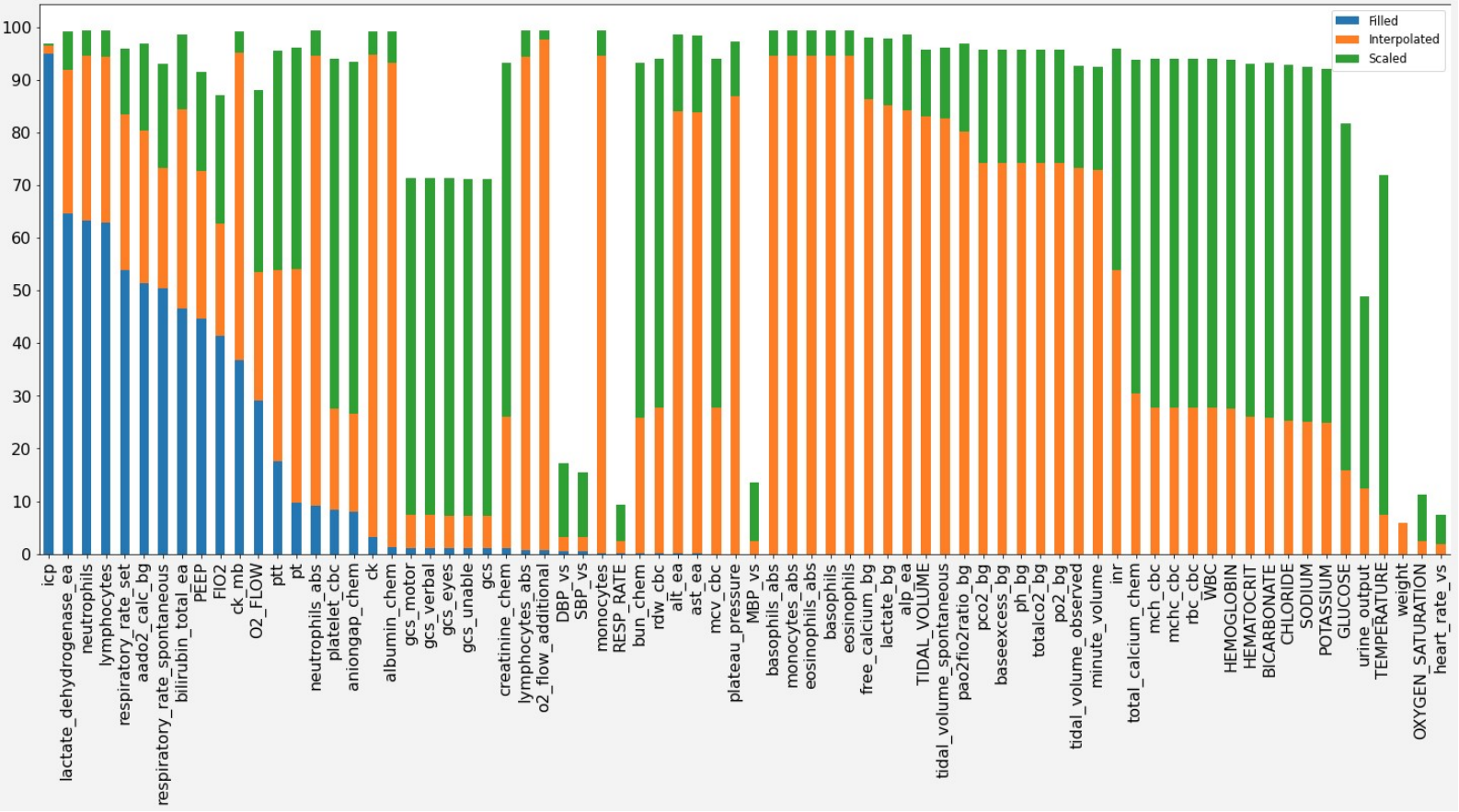


Figure 4.3: Shows the change in percentage missingness of each feature with successive data imputation techniques. The green bar indicates the rate of missingness of the original dataset. After step 2 of the imputation methods, the missingness of each feature is reduced to the orange level. Finally, with the last stratification and imputation step, we are left with missingness levels indicated by the blue bars on the bar chart.

Vital signs had the most complete measurements, with fewer than 15% of values missing, which was further reduced to below 5% after imputation. The blood gases biomarker, however, exhibited a higher level of missingness, ranging from 94.4% to 99.3%. After imputation, the missingness decreased to less than 50%.

Laboratory results generally had the highest proportion of missing data, likely due to the selective nature of testing in critical care. Laboratory tests are typically ordered based on the patient's immediate clinical needs, with only certain tests being prioritised at any time. As a result, many laboratory measures may not be consistently recorded for all patients, contributing to higher missingness. After imputation, the missingness for most laboratory measures fell to below 5% (Figure 4.3).

The Glasgow Comma Scale (GCS) and its components are typically measured upon admission and at 4-hour intervals. This rate of assessing GCS resulted in 75% missingness of the data. Nonetheless, the proportion of GCS missingness fell to 10% after interpolation and decreased to under 5% after forward-filling.

Overall, interpolation has had the highest impact on reducing missingness, except for the blood gases measure, in which the highest impact was attributed to the final imputation step.

4.5.3 Model performance

Diag. point	Dataset	Full Set of Features				N. Feat Selected	Reduced Set of Features			
		Accuracy	AUROC	Sensitivity	Specificity		Accuracy	AUROC	Sensitivity	Specificity
t_{sepsis}	1	89.91%	89.05%	84.10%	94.01%	16	75.02%	72.87%	60.41%	85.33%
	2	95.80%	95.18%	91.61%	98.76%	16	84.00%	81.92%	69.91%	93.93%
	3	**99.40%**	**99.35%**	**99.07%**	**99.62%**	12	93.14%	92.22%	86.88%	97.56%
t_{sepsis} - 6	1	93.29%	92.89%	89.94%	95.84%	13	73.06%	71.50%	60.11%	82.90%
	2	95.57%	95.08%	91.50%	98.65%	15	82.47%	80.70%	67.79%	93.60%
	3	98.88%	98.78%	98.04%	99.53%	16	**93.08%**	**92.36%**	**87.04%**	**97.67%**
t_{sepsis} - 12	1	93.69%	93.37%	90.59%	96.14%	13	72.60%	71.33%	60.45%	82.20%
	2	95.46%	95.06%	91.50%	98.62%	16	81.55%	80.08%	67.26%	92.89%
	3	98.86%	98.77%	98.01%	99.53%	14	91.90%	91.19%	85.21%	97.17%

Table 4.3: Model Performance across diagnosis points, point 0 (sepsis-3 diagnosis time), 6 (6 hours before point 0) and 12 (12 hours before point 0) and for full and reduced feature sets. The best performances are highlighted.

As expected, more levels of imputation resulted in better diagnostic performance (Table 4.3). The average AUROC for models trained on non-imputed data, models with interpolation and models with the addition of forward-filling/median imputation were 91.8%, 95.1% and 99.0%, respectively. Adding levels of imputation improved accuracy by 10% for diagnosing sepsis at the original time. Models trained on the complete feature set that have undergone all three

stages of imputation performed best at predicting sepsis at t_{sepsis} (99.4% accuracy), with performance decreasing by only 0.6% when using earlier diagnosis times.

The AUROC of the worst model was 71.3%, which was trained on non-imputed data with a reduced set of features and a diagnosis point set to 12 hours before sepsis-3 diagnosis. The same feature-reduced dataset at t_{sepsis} had a 17.7% increase in Area Under the Receiver Operating Characteristic curve (AUROC).

Each successive imputation step improved the diagnostic performance in feature-reduced models by an average of 10% in AUROC per step. The model trained on the complete set of features with no imputation achieved 89.1% AUROC, similar to other studies using the same diagnostic criteria [150–152, 158].

Emphasis was placed on evaluating the model's performance at the earliest possible diagnosis point, t_{onset}. In rightly classifying cases at t_{onset}, the model achieved an AUROC of 99.53% using data that has undergone all stages of modification. This increased slightly by 0.04% when the diagnosis point was set to $t_{suspicion}$.

Overall, ranked in the order of highest sensitivity and specificity, the best performances were achieved with suspected infection, followed by onset and finally Sepsis-3. Performance of evaluating the model at the different clinically relevant diagnosis points and stages of imputation is recorded in Appendix Table A.2.

The following section explores the features identified within the models used in this study.

4.5.4 Correlation and Feature Importance

Figure 4.4 presents the correlation matrix for the top 12 features at the t_{sepsis} diagnosis point. The correlations ranged from -0.93 to 0.66, reflecting a range of relationships between these clinical variables. The following analysis focuses on the top five correlations, which provide critical insights into sepsis management and patient prognosis.

The strongest negative correlation observed was between GCS verbal scores and ICU length of stay (LOS), with a coefficient of -0.93. This suggests that patients with lower GCS verbal scores, which indicate more severe neurological impairment, tend to have longer ICU stays. This inverse relationship highlights the intensive care needs of neurologically compromised patients, who often require prolonged respiratory support, continuous monitoring, and management of multiple organ systems, resulting in their ICU stay being longer.

A significant positive correlation (0.69) was found between GCS verbal scores and invasive ventilation. This indicates that patients with lower GCS verbal scores are more likely to

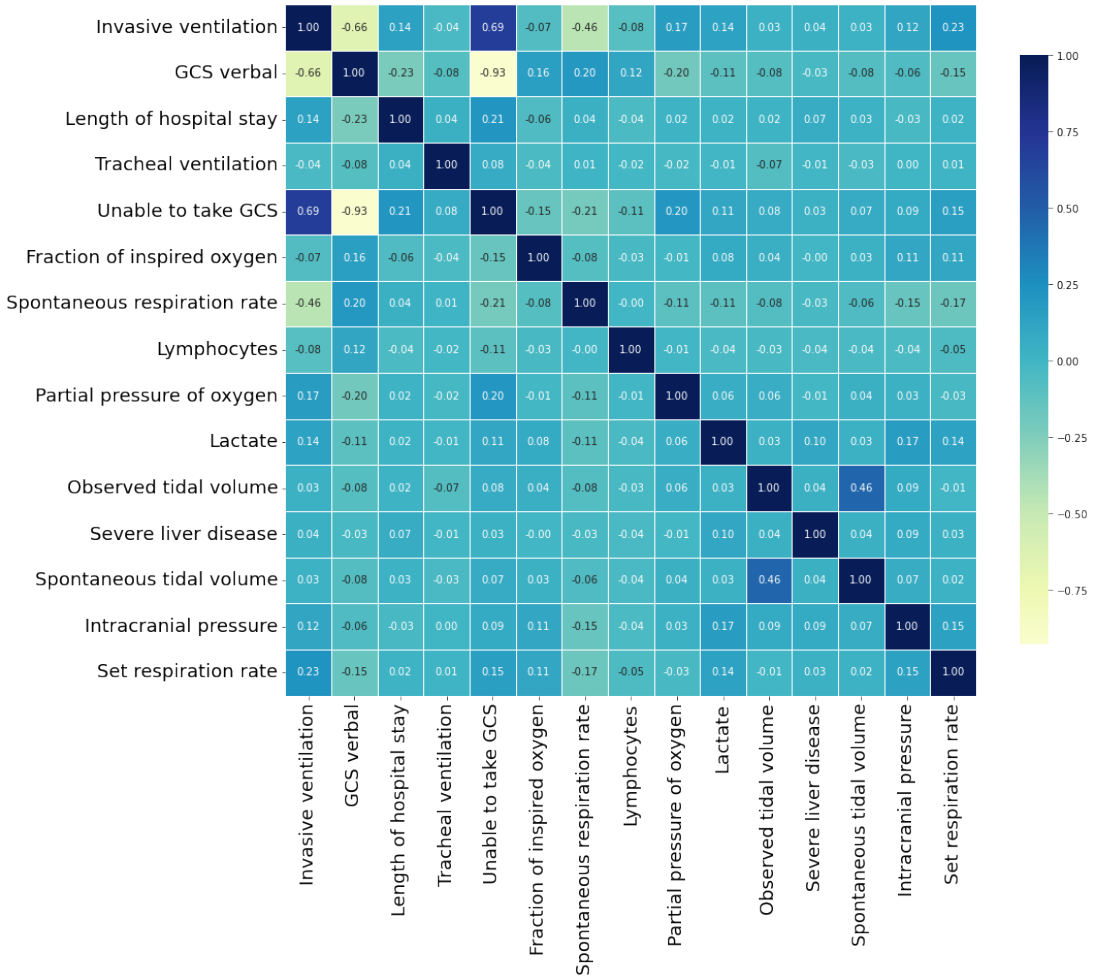


Figure 4.4: Correlation plot of the top 12 features.

require invasive ventilation, as compromised neurological function often necessitates mechanical ventilation. However, this relationship also points to a clinical challenge: intubated patients usually cannot respond verbally, which could artificially lower their GCS verbal scores. This presents a difficulty in assessing neurological status in ventilated patients, further complicating the use of GCS verbal scores as a reliable measure in such contexts.

The correlation between Fraction of Inspired Oxygen (FiO_2) and spontaneous respiration rate (0.67) suggests a compensatory mechanism in critically ill patients. As the need for oxygen increases (reflected by higher FiO_2 levels), the spontaneous respiration rate also rises, likely as the body attempts to improve oxygenation. This correlation underscores the dynamic nature of respiratory response to critical illness and the interplay between oxygen delivery and

spontaneous breathing efforts.

The relationship between FiO_2 and observed tidal volume (0.54) further emphasises the interdependency of respiratory parameters. As FiO_2 increases, adjustments to tidal volume are often necessary to ensure that the patient receives adequate ventilation. This highlights the importance of carefully managing oxygen levels and ventilation volume, as optimising these parameters is essential for maintaining patient stability, particularly in patients with respiratory compromise.

A moderate positive correlation (0.47) between lactate levels and observed tidal volume suggests that metabolic acidosis, typically caused by tissue hypoxia or ischaemia, may influence ventilatory settings. Higher lactate levels indicate impaired tissue perfusion, which can necessitate adjustments in ventilation, including tidal volume, to improve oxygenation and acid-base balance. This relationship highlights the interconnectedness of metabolic and respiratory responses in critically ill patients.

These correlations were also reflected in the feature importance analysis. GCS verbal scores and invasive ventilation status were integral in predicting sepsis in our models. Models trained on non-imputed data tended to prioritise invasive ventilation, GCS verbal scores, and ICU length of stay, underscoring their importance in sepsis prediction. After imputation, however, the focus shifted to variables such as FiO_2 and GCS verbal, aligning more closely with clinically relevant diagnostic criteria. The final imputation step further highlighted GCS verbal as a key diagnostic biomarker and emphasised other vital markers, including lactate, respiration rate, and partial pressure of oxygen (Figure 4.6).

4.6 Contextualising clinical findings

With the availability of Electronic Health Records (EHR) and advancements in modelling approaches over recent years, many have identified key biomarkers for accurate sepsis diagnosis and created opportunities for early diagnosis using complex algorithms. However, without appreciating the nuances in missingness in healthcare data, some have oversimplified the methods of handling missing data during the data preparation stage.

Missing values in clinical data, such as the MIMIC-IV dataset, are not arbitrary and should be handled meticulously. Studies applying generic data imputation methods (e.g., [153, 159]) have limited applicability in clinical settings. Hence, our work applies clinically informed methods of handling data that account for the complexities of ICU data and evaluate model performance compared to similar studies.

Feature	0 hours before			6 hours before			12 hours before		
	Scaled & Encoded (1)	Interpolated (2)	Forward Fill or Median (3)	Scaled & Encoded (1)	Interpolated (2)	Forward Fill or Median (3)	Scaled & Encoded (1)	Interpolated (2)	Forward Fill or Median (3)
Invasive ventilation	0.0521		0.0031	0.0526		0.0037	0.0532		0.0036
GCS verbal	0.0337	0.0377	0.0636	0.0356	0.0398	0.0527	0.0297	0.0410	0.0605
Length of hospital stay	0.0281	0.0206	0.0131	0.0234	0.0189	0.0149	0.0221	0.0189	0.0140
Tracheostomy ventilation	0.0230		0.0017	0.0242		0.0152	0.0237		0.0042
Unable to obtain GCS	0.0157	0.0038	0.0020	0.0051	0.0044	0.0021	0.0149	0.0052	0.0033
Fraction of inspired oxygen	0.0156	0.1409	0.0199	0.0155	0.1389	0.0151	0.0163	0.1514	0.0164
Spontaneous respiration rate	0.0013	0.0253	0.0307	0.0012	0.0114	0.0330	0.0011	0.0159	0.0290
Lymphocytes	0.0011	0.0235	0.0082	0.0021	0.0235	0.0096	0.0022	0.0230	0.0097
Oxygen pressure in the blood	0.0018	0.0199	0.0507	0.0014	0.0201	0.0218	0.0012	0.0133	0.0216
Lactate	0.0011	0.0065	0.0298	0.0014	0.0065	0.0277	0.0014	0.0059	0.0262
Observed tidal volume	0.0013	0.0040	0.0216	0.0014	0.0049	0.0195	0.0012	0.0053	0.0186
Severe liver disease	0.0152	0.0179	0.0116	0.0174	0.0184	0.0153	0.0150	0.0199	0.0143
Spontaneous tidal volume	0.0026	0.0153	0.0057	0.0016	0.0315	0.0059	0.0022	0.0133	0.0058
Inter-Cranial Pressure (icp)	0.0106	0.0174	0.0221	0.0105	0.0169	0.0194	0.0105	0.0186	0.0211
Set respiration rate	0.0024	0.0075	0.0095	0.0018	0.0068	0.0442	0.0017	0.0077	0.0438

Figure 4.5: Feature importance scores for the top 15 features at 0, 6, and 12 hours before sepsis-3 diagnosis. Darker shades indicate higher importance in predicting sepsis.

The baseline model trained on the complete feature set was compared before and after the elaborate data preprocessing steps were applied. Overall, data imputation steps improved sepsis detection at t_{sepsis}. The AUROC, specificity, and sensitivity respectively increased from 89.05 to 99.35%, 94.01% to 99.62%, and 84.10% to 99.07%. The successive imputation steps were significant, transitioning the model from a diagnosis relying heavily on invasive ventilator status to one that assessed the patient's verbal response and oxygenation levels (as shown in Table 4.5). The models performed well in diagnosing sepsis 6 and 12 hours before, suggesting that early sepsis diagnosis is possible. The highest AUROC achieved was 99.35% when the diagnosis point was set by Sepsis-3 criteria, outperforming similar studies.

The high importance of artificial ventilation can be attributed to the cohort's average age of 63.7 years. This is consistent with the findings of Farzan et al. [175], who found that older patients were twice as likely to require invasive ventilation. Beyond age, the use of ventilation indicates a patient's deteriorating condition. The use of a ventilator suggests that the patient may be on the verge of respiratory failure or suffering from severe hypoxaemia, which are known manifestations of sepsis.

A key aspect of this study's novelty is the decomposition of the Glasgow Coma Scale (GCS), which is typically aggregated in sepsis models and diagnostic tools. This research is the first to decompose GCS into its components. By doing so, the model found that GCS verbal scores were a key prognostic marker for sepsis. This granular approach offers a more precise prediction than previous models that treated GCS as an aggregated score.

Further, the study found that the FiO_2 and Partial Pressure of Oxygen (PaO_2) were critical

in assessing the patient's susceptibility to hypoxaemia, a key early sign of sepsis. Hypoxaemia results from the hypoperfusion of oxygen in the blood and can lead to long-term organ damage and potential mortality if left untreated [176]. Hypoxaemia in the brain often results in an impaired level of consciousness, contributing to sepsis-associated delirium. Patients with delirium often struggle to follow verbal commands and exhibit slurred or confused speech [177, 178]. These observations further validate the importance of monitoring the verbal component of the GCS as a potential early indicator of sepsis.

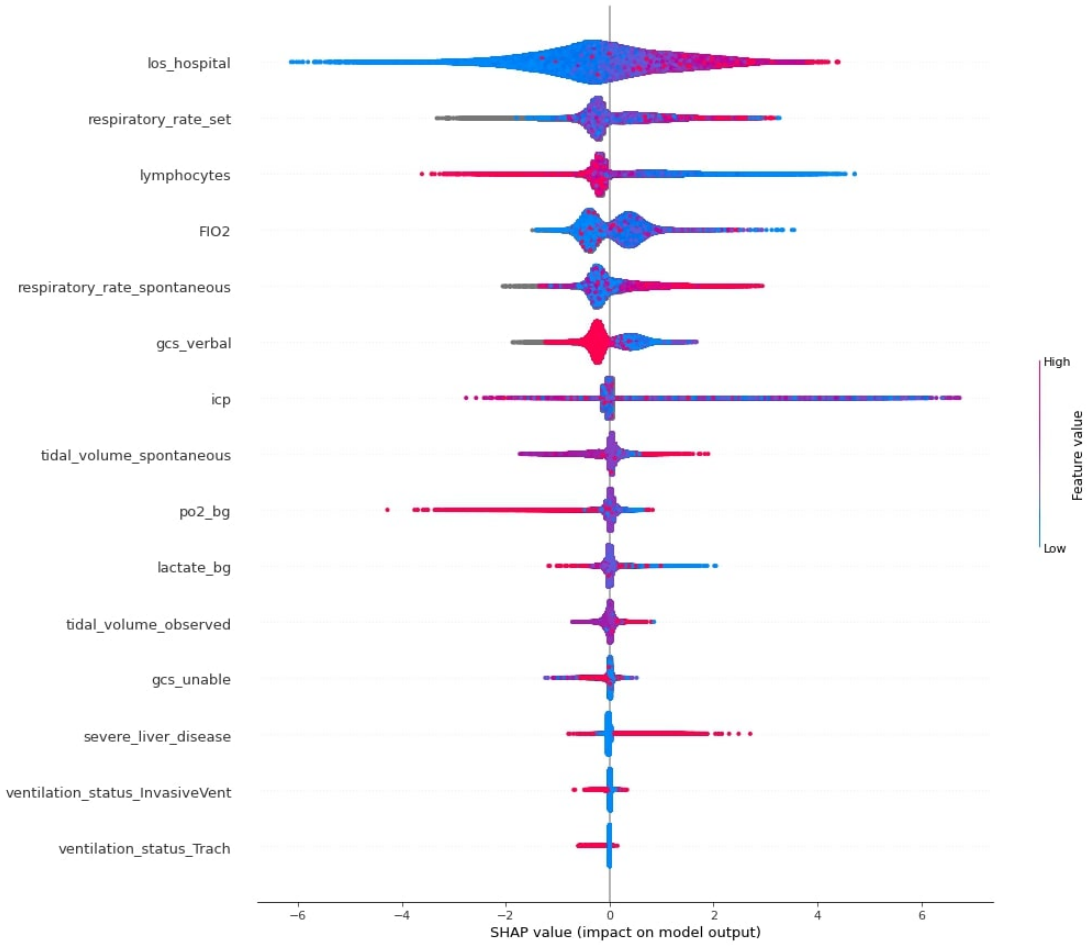


Figure 4.6: Violin plot that offers insight into the distribution of values in each feature and its impact on the patient's assessment. The hue axis indicates high values recorded for the specific feature. Positive impact indicates that patients are more likely to be diagnosed with sepsis, and vice versa.

Tree SHAP was used to ensure the validity of feature importance metrics. These values provided profound insights into the effect of key features on the model's prediction. As shown

in Figure 4.6, the distribution of the most important features, represented in red and blue as high and low values, validated the significance of GCS verbal as a prognostic marker. Low GCS verbal scores were correlated with a higher likelihood of a sepsis diagnosis (positive SHAP values). Additionally, the model identified the impact of other vital biomarkers, suggesting that extended hospital stays before ICU admission, high respiration rates, and tidal volume contributed to diagnosing sepsis [176]. Interestingly, low lymphopenia was another indicator that influenced the model's decision, corroborating findings in Drewry et al. (2014) that low lymphocyte counts are linked to sepsis mortality [179].

Beyond the other predictors, this study reinforces that early detection of sepsis can be achieved by observing changes in the patient's level of consciousness, which can occur well before invasive interventions like ventilation are necessary. Early diagnosis could translate to reduced mortality rates, higher sepsis survival rates, and fewer patients experiencing the long-term sequelae of severe sepsis. Further clinical research and analysis are needed to substantiate these findings. Moreover, the methods for handling missing data employed in this study could serve as a framework for future research in healthcare data, once the findings are validated further.

4.6.1 Limitations/Assumptions

The MIMIC dataset used in this study only includes data from one hospital, which means the findings may be more relevant to the population demography specific to that institution. Thus, the results may not be fully generalisable to other clinical settings or populations. The study's scope is limited to ICU patients, so the conclusions may not extend to patients in other healthcare environments.

While widely accepted, the Sepsis-3 criteria used for diagnosis may have limitations in cases where patients have a shorter ICU stay, potentially affecting diagnostic accuracy. Therefore, the findings in this study are dependent on the reliability of the Sepsis-3 criteria itself, which has been noted in the literature for its reduced diagnostic performance in short-term ICU patients [137].

Key assumptions in this study include:

- The admission SOFA score was assumed to be based on the biomarker levels from the hospital stay before ICU admission. For patients without prior hospitalisation, the SOFA score was derived using vital signs taken upon ICU admission.

- The importance metrics derived from the machine learning (ML) models were assumed to accurately reflect the clinical markers that have the most significant impact on the early diagnosis of sepsis.

4.7 Summary

This study aimed to answer two key questions: 1) How does a clinically informed model perform compared to existing models in sepsis detection, and 2) Are the most impactful features aligned with our clinical understanding of sepsis? By leveraging advanced data processing methods, specifically clinically informed handling of missing data, the models achieved superior performance in sepsis detection compared to existing models, enabling early detection up to 12 hours before the Sepsis-3 criteria diagnosis point. Models trained on the complete feature set outperformed those using reduced features by at least 20% in AUROC. This demonstrates that sepsis diagnosis requires a comprehensive set of features, even for real-time or advanced detection.

One of the key innovations in this work lies in the clinically informed preprocessing pipeline for ICU markers, which proved essential for enhancing the predictive power of the models. Furthermore, by decomposing the GCS into its components, the study highlighted the importance of the GCS verbal score as a key prognostic marker in sepsis detection. This approach, which contrasts with traditional aggregated GCS models, provided valuable insights into early sepsis prediction.

Additionally, the study employed TreeSHAP to ensure the interpretability and explainability of the models. This makes the results more accessible to clinicians, addressing concerns about using complex ML tools in critical care settings. The models identified key indicators of sepsis, such as multiple hospitalisations before ICU admission, invasive ventilation, and GCS verbal score. While existing literature supports the relevance of these features, the study further establishes their importance in sepsis detection and highlights the need for further exploration of these associations.

The study successfully applied domain expertise to improve data handling techniques, developed highly accurate diagnostic models, and made significant strides in understanding the importance of early biomarkers like GCS verbal scores and FiO_2. These contributions lay the foundation for future work applying AI in clinical settings, particularly improving sepsis outcomes in ICU.

4.8 Future Work

Future work will focus on validating GCS verbal and ventilation settings in other ethnic groups, including minorities, and obtaining databases from a range of hospital ICUs.

Additionally, further interpretability techniques and other ML techniques can be explored to validate the importance and generalisability of the identified key diagnostic biomarkers. This may lead to exploring the use of SHAP as a real-time interpretability tool for clinicians to utilise in making informed decisions in ICUs. This will require user trials on clinicians using these tools in fast-paced environments and iterative improvements to ensure the tool is optimised for real-time, clinical settings.

4.9 Connecting the dots and remembering the why

The primary focus of this research has been to improve the interpretability and reliability of ML models designed to predict sepsis at an early stage during an ICU stay. This endeavour is motivated by the urgent need to improve sepsis outcomes and the broader goal of advancing the integration of ML in healthcare. Effective and early sepsis detection can significantly improve patient prognosis, reduce mortality rates, and alleviate the burden on healthcare systems. This work contributes to the growing evidence that AI and ML models can play a crucial role in clinical decision support, particularly in high-stakes environments like the ICU.

Chapter 5

The Emergence of COVID-19: Mining Temporal Patterns

Contents

5.1 A World caught off guard

The year 2020 will forever be marked in history as the year the world was caught off guard by the emergence of a novel coronavirus, SARS-CoV-2, causing the disease known as COVID-19. Starting from Wuhan, China, in late 2019, the virus quickly spread across the globe, leading the World Health Organisation to declare a pandemic in March 2020 [19]. The virus's rapid spread, its high transmission rate and the severity of symptoms in many cases, resulted in COVID-19 quickly becoming an unprecedented global health crisis. Countries worldwide scrambled to respond, implementing measures ranging from travel restrictions and lockdowns to mass testing and contact tracing efforts. Despite these efforts, the virus continued to spread, overwhelming healthcare systems and leading to a significant number of deaths [19].

While the COVID-19 pandemic has affected millions of people worldwide [180], hospitalisations due to severe COVID-19 cases can be considered a rare event, particularly in the early stages of the pandemic. Most individuals infected with the SARS-CoV-2 virus experienced mild to moderate symptoms and did not require hospital care [181]. However, the small percentage of cases that did require hospitalisation had a profound impact on healthcare systems, leading to overwhelming strain on resources and, tragically, a significant number of deaths. Investigating the early manifestations of COVID-19, including signs and symptoms that could be identified in primary care settings, became crucial in predicting which patients were at higher risk of developing severe complications. By focusing on these early indicators, healthcare professionals aimed to identify and closely monitor high-risk patients, facilitating timely interventions and resource allocation. Predicting and managing these rare but consequential events through recognising early warning signs in primary care became a critical challenge in the fight against COVID-19.

The pandemic exposed the vulnerabilities of even the most robust healthcare systems, highlighting the need for improved disease surveillance, early detection, and rapid response capabilities [182]. The emergence of COVID-19 served as a stark reminder of the ever-present

threat of emerging infectious diseases and the need for ongoing research and preparedness to respond to such threats effectively.

5.2 The first line of Defence

As the pandemic unfolded, GPs found themselves at the forefront of the battle against COVID-19. GPs play a crucial role in the healthcare system, often serving as the first point of contact for patients [183]. In the context of the COVID-19 pandemic, their role became even more critical in identifying and managing potential cases while ensuring continuity of care for patients with chronic conditions and other health concerns.

They had to balance the need to protect these patients from potential exposure to the virus while ensuring they received the care and support needed [183]. Moreover, GPs were instrumental in recognising early warning signs and identifying patients at higher risk of developing severe complications, enabling timely interventions and referrals to specialist care when necessary.

As the virus spread and hospitals became overwhelmed, GPs were tasked with triaging patients by determining who needed immediate hospital care and who could be safely managed at home [184]. They had to quickly adapt to new ways of working, with many consultations moving online or over the phone to reduce the risk of transmission. Further, GPs played an instrumental role in the rollout of the COVID-19 vaccines by administering them and providing information and reassurance to patients [183].

5.3 The advent of Self-Attention

The advent of self-attention mechanisms, particularly in transformer models, has revolutionised the field of ML [185]. These mechanisms allow models to focus on different parts of the input sequence when making predictions, providing a form of context-awareness that was previously difficult to achieve [186].

In the context of the COVID-19 pandemic, self-attention mechanisms have shown great promise in various applications. For instance, they have been used to analyse temporal patterns in the spread of the virus [187], helping to predict future outbreaks and inform public health responses.

Self-attention mechanisms have also been used to analyse EHR, helping to identify patients at high risk of severe outcomes from COVID-19 [188]. By focusing on relevant parts of a

patient's history, these models can make more accurate predictions, helping to guide clinical decision-making.

Moreover, the interpretability of self-attention mechanisms makes them particularly useful in healthcare settings. By visualising the attention scores, clinicians can gain insights into the model's decision-making process, which could help establish trust in the model's decisions.

5.4 Current Challenges and Opportunities

Despite COVID-19 prevention and risk mitigation efforts, hospital resource utilisation remains a significant concern. Although the hospitalisation rate has reduced from 36.7 to 8.2 per 100,000, UKs healthcare system still faces 25% of the strain felt during the first wave [189]. With the resurgence of COVID-19 and its variants, patients at risk of hospitalisation due to the virus must be rapidly identified for early intervention and risk mitigation.

Numerous studies developed AI predictive tools to identify patients at risk of severe outcomes due to COVID-19. These tools utilise highly interpretable ML algorithms, such as decision trees [190–192], gradient boosting decision trees [193–195] and logistic regression [196, 197]. These models yielded Area Under the Curve (AUC) scores between 0.74 and 0.92 and discovered critical prognostic markers essential for predicting a patient's COVID-19 outcome, including white cell differential count, creatinine phosphate and lymphocyte proportion.

However, identifying risk factors leading to COVID-19 hospitalisation proved difficult for ML models, primarily when large populations and multiple comorbidities are involved. Willette *et al.* used a permutation-based linear discriminant analysis to predict COVID-19 and hospitalisation risk [198]. When trained on a subset of participants with an antibody titer, their models achieved an AUC of 0.969 (95% CI 0.934–1.000), but when trained on a more significant portion of the population, the AUC dropped to 0.803 (95% CI 0.663–0.943). Similarly, Wollenstein *et al.* [196] only achieved 61% accuracy predicting COVID-19 hospitalisations.

RNNs are known for modelling long temporal sequences in heterogeneous patient data [199–201]. Besides a lack of interpretability, DL models like RNNs are limited in handling dimensionally varying or missing data. These constraints have led to using 1D Convolutional Neural Networks (1D-CNNs), which generates feature importance similar to logistic regression models. 1D-CNNs cannot model temporal patterns well and often underperform RNNs.

This study aims to employ the capabilities of RNNs in predicting COVID-19 hospitalisations whilst providing interpretability and improving on REverse Time AttentIoN model (Retain), an interpretable RNN [200] and its successor REverse Time AttentIoN EX model

(RetainEX) [201]. Retain and RetainEX adopt a temporal attention generation mechanism to learn the importance of each GP visit and each medical code. However, they lack architectural depth, visit-level risk scores and the ability to learn from imbalanced datasets.

Therefore, this study aims to improve the predictability and interpretability of the Retain models and demonstrate an enhanced version of the model, RetainEXT. Study findings may aid in identifying individuals with a high risk of hospitalisation from the virus, thereby allowing for early interventions to mitigate risks of post-infection complications. This study hopes to show the feasibility of applying interpretability methods to RNN models, hoping to encourage their use in similar rare event detection algorithms.

This chapter begins with the study design and population and a description of RetainEXT in Section 5.5, followed by evaluating and comparing the results and feature importance extrapolated from the best-performing RetainEXT model in Section 5.6. Finally, it concludes and elaborates on future works in Section 5.7.

5.5 Integrating advanced approaches to hospitalisation risk modelling

5.5.1 Study Design

The study design was a retrospective longitudinal Self-Controlled Case Series (SCCS). Data sources used were the Welsh Demographic Service Data (WDSD), primary care GP data using the Welsh Longitudinal General Practice (WLGP) dataset and secondary care data using the Patient Episode Data for Wales (PEDW) dataset, all held within the Secure Anonymised Information Linkage databank [1iteLyons2009. The SAIL databank contains over EHR of 80% of the Welsh primary care data, making it an ideal data resource. Analysis was conducted within SAIL's Research Environment. We followed the patient's history between 2009 and 2020 from their earliest interactions with the NHS up to and including their first COVID-19-related hospitalisation.

Patients were included if they had 1) a minimum of 2 GP interactions in the data collection period, 2) were aged 18 and above at the start of the data collection period, and 3) had a hospital admission where, within 14 days of admission, they received a positive COVID-19 test taken during the hospitalisation episode.

To limit confounding factors, we excluded medical codes, including COVID-19 infection or hospitalisation, collected within the 14 days before a positive test result.

5.5.2 Data Structure

Similar to the structures used by Choi *et al.* [200], we modelled each patient's EHR as Encounter Sequence Modelling (ESM) [199], where the sequence of patient visits is represented by a varying number of medical codes $d^1, ..., d^l$, where l is the number of diagnosis codes per GP visit. Here, x_t represents the visit, and d^j is the j^{th} code from the dictionary of all codes, D. Therefore, the total number of possible codes is $r = |D|$.

Traditionally, ESM models each visit $x_t \in \{0,1\}^{|D|}$ as a binary vector, where the value 1 in the j^{th} coordinate indicates that d^j was documented in the t^{th} visit. Given a sequence of visits $x_1, ..., x_T$, with T as the total number of visits, the goal of ESM is to predict the codes occurring at the following visit $x_2, ..., x_{T+1}$, with the number of labels $y = |D|$.

As $|D|$ contains 27,317 unique read codes, it would be resource-heavy and impractical to train a model on large binary vectors. Instead, we decided that a patient's record, x_t, should only include the medical codes recorded during that visit.

Raw medical codes contain a mixture of alphanumeric characters; therefore, we decided to encode $|D|$ into sequential numerical values with arbitrary meaning. Having defined each patient's visit as $x_t = d_1, ..., d_s$, we set the model to predict the patient's risk of hospitalisation, $\hat{y_t}$, at each visit, x_t.

Following each visit, x_t was passed through an embedding layer to learn the representation and later visualise any clusters of medical codes. This generates v_t for each x_t, then concatenates to the patient's age and gender. Both variables were included at every time step due to the patient's varying age and possibly gender.

Furthermore, the model's comprehension of the time between visits is vital to determining the risk of hospitalisation at each visit. For instance, a series of visits to the GP over a short period may indicate comorbidity or severe illness. Long hibernation may suggest good health and influence the model to predict lower risk scores. To harness temporal information, we incorporate visit dates as an additional feature.

Given a sequence of T events $t_1, t_2, ..., t_T$, we obtain $\Delta t_i = t_i - t_{i-1}$ for each successive visit. We assume that the first visit is unaffected by time constraints by fixing Δt_1 to 1. We explored the benefit of additional representations of time, which are (1) Δt_i the time interval between visits) [200], (2) $1/\Delta t_i$ (its reciprocal value) [202], and (3) $1/\log e + \Delta t_i$ (an exponentially decaying value) [202]. These values are concatenated to the embedding v_t for each x_t, to enrich the information for our model.

While handling the data, we found a critical improvement necessary to improve the epi-

demiological study design of Retain and RetainEX.

5.5.2.1 Per-event risk assessment

Retain and RetainEX both utilise an Learn to Diagnose (L2D) [203] approach and are limited to predicting a risk score for each patient. This suggests that both case and control groups were used to learn high-risk markers, hindering clinicians from evaluating outcome severity at each GP visit. Thus, our study modified the algorithm's output to model a risk score at each GP visit using an SCCS study design.

The SCCS is a case-only method in which confounders are automatically controlled for [204]. This allows us to investigate the association between a transient exposure and an outcome event, helping clinicians make better-informed decisions.

5.5.3 RetainEXT

Fig. 5.1 (A) shows our model takes in a patient visit sequence as C dimensional vectors $x^1, x^2, ..., x^T$. An embedding matrix $W_{emb} \in R^{m \times C}$ is used to convert all 27,317 unique medical codes linearly into a matrix of size m × C, where m is the number of units in the embedding layer, resulting in $v_t = W_{emb} \cdot x_t$. The patient's age and gender, n_t, are also appended to v_t at each visit and passed through a dropout layer to improve the model's generalisability. Additionally, each successive visit generates a set of representations that characterise the time between visits, Δt^t, to offer additional insight into the patient's state; this is concatenated to v_t and n_t.

Following the structure of Retain and RetainEX, we computed two attention types, α and β. Fig. 5.1 (B) and (C) represent the stacked Bi-LSTM network that takes in the age, gender and time-attached visit representations and returns attention values (e.g., contribution scores).

α_t is a single value representing the importance of each GP visit. β_t is an m-dimensional vector that quantifies the significance of each medical code within a specific visit. To benefit from both visit and feature-level importance requires a separate stacked Bi-LSTM to compute each attention class.

For each $[v_t; n_t; \Delta t^t]$, the stacked α-Bi-LSTM computes the forward and backward hidden states of the first α—Bi-LSTM, then passes it on to the second α-Bi-LSTM. The final hidden state vectors, g_t^{f2} and g_t^{b2}, are concatenated into a single m-dimensional vector, which is passed on to a dense layer.

The parameter $w_\alpha \in R^{2m}$ was used to compute a scalar value for each time step as $e_t = w_\alpha [g_t^{f2}; g_t^{b2}]$. Next, the softmax function is applied to all scalar values $\{e^1, ..., e^T\}$ to obtain

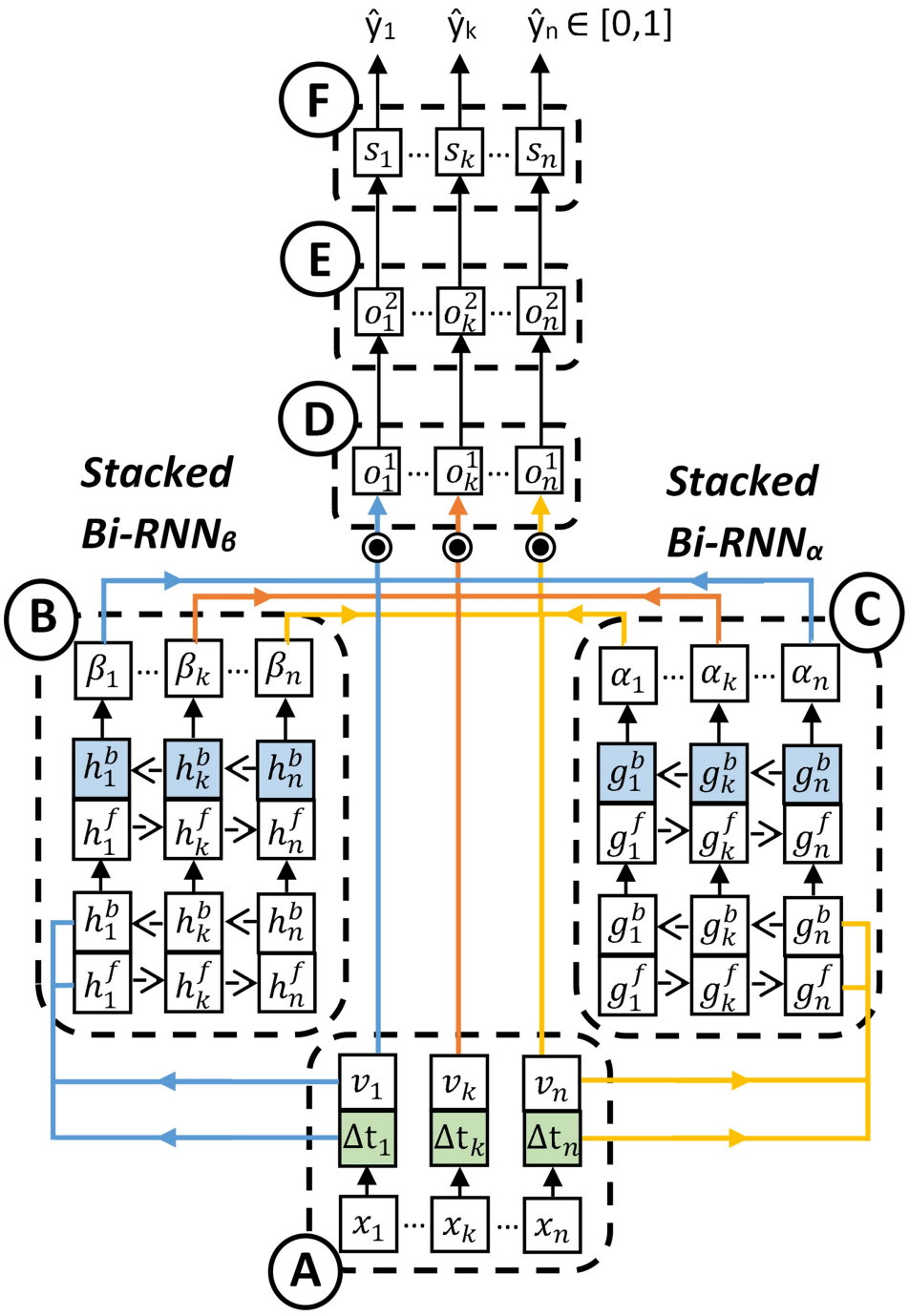


Figure 5.1: Overview of RetainEXT. (A) Using a single embedding layer, a binary vector x_t is represented as embedding vectors v_t, with time interval information appended to the former. (B, C) v_t is input into two Bi-LSTM layers to obtain scalar α and vector β attention weights. (D) α, β and v_t are multiplied over all time steps, and then each time step is passed through a dense layer. (E) Output from the first dense layer is dimensionally reduced into a single output per time step. (F) Each time step output is non-linearly transformed to a risk score $\hat{y}$.

$\{\alpha_1, \alpha_2, ..., \alpha_T\}$, a distribution of attention values that sum to one. Similarly, the concatenated hidden state vectors generated by the stacked β-Bi-LSTM are multiplied by $w_\beta \in R^{mx2m}$ and return an m-dimensional vector β_t for the t^{th} visit as $\beta_t = w_\beta\ [g_t^{f2}; g_t^{b2}]$.

After obtaining both α_t and β_t values, we performed element-wise multiplication with the concatenated array $[v_t; n_t; \Delta t^t]$, and passed it through a dense layer with weights w_o^1 (Fig. 5.1D). The output is then passed to the final dense layer with weights w_o^2. The additional dense layer increased the complexity of providing interpretability; thus, we defined w_{out}^t as each event's aggregated and combined weight. In mathematical terms, $w_{out}^t = \Sigma_z^U \Sigma_g^T w_{o,g}^1 \cdot w_{o,z}^2$, where U is the number of units in the 2^{nd} dense layer.

Lastly, the contribution score for each visit was computed, $s_t = w_{out}^t \cdot o_t^1$. The t^{th} visit is passed through a dense layer with sigmoid activation to dimensionally reduce the vector at each time step, allowing us to compute a normalised prediction value, $\hat{y}_t$, ranging between 0 and 1 where $w_{out}^t \in R^m$. The predicted value indicates the patient's risk of hospitalisation in that particular visit, with a value closer to 1 indicating a higher risk. We trained our model to minimise binary cross-entropy loss and conducted hyperparameter tuning for a fraction of dropout, regularisation, embedding and stacked LSTM units. Here, we found improvements needed to limit the input data to only relevant medical codes and improve performance, especially when modelling rare events.

5.5.3.1 Stacked and Deeper Architecture

The original implementations of Retain and RetainEX lacked depth in their architectures. Stacked Bi-LSTM have increased the capacity to identify complex nested patterns in patients that single-layer networks may overlook. Using an additional dense layer before the output layer also facilitates understanding non-linear patterns in the dataset. However, the added layers increase training time. Section 5.6 compares the models' performance gain and training times before and after including the additional layers.

5.5.3.2 Ragged Tensors

Another novelty of this study is the adoption of ragged tensors to compute patients with a variable number of visits and medical codes in each visit. This eliminates the need for padding and masking or binary encoding of medical codes, reduces training time, and improves model performance.

5.5.3.3 Rare event weighting

RetainEXT implements the use of sample weighting, which not only attributes a single weight to a patient but also at each visit. Hospitalisation is considered a rare event as only 1.59% GP visits of the entire dataset of medical codes are related to hospitalisation due to COVID-19. Thus, rare event weighting helps the model learn significantly more from a few samples. In contrast, the preceding Retain and RetainEX models struggle to model an imbalanced scenario.

5.5.4 Interpretability

The backbone of Retain and RetainEX is their long-standing ability to provide feature- and visit-level importance scores. With this foundation, we focused on understanding the global feature importance, which is an information-rich measure of how clinically aligned our model is.

5.5.4.1 Local attention

RetainEXT and its predecessors achieve its transparency by multiplying the final layer of the stacked Bi-LSTM generated attention weights α_t and β_t to the visit vectors v_t to obtain the context vector o_t^1, which are used instead of the Bi-LSTM hidden state vectors to make predictions. Each input vector x_t has a linear relationship with the final contribution score, S. Thus, we derive an equation that measures the contribution score of the code d at time step t to S by reformulating the equations above as $S_t^d = \alpha_t w_{out}(W_{emb}[d,:] \cdot \beta_t)$, where $W_{emb}[d,:]$ is the d^{th} row of W_{emb}.

Additionally, we generate a visit-level contribution score S_t by aggregating contribution scores of codes for each visit as $S_t = \Sigma_{d \in x_t} S_t^d$.

5.5.4.2 Global feature importance

Retain and RetainEX perform dimensionality reduction on the embedding weight of all clinical markers to visualise possible clusters in the data. However, Kwon *et al.* [201] mentioned that limitations exist in visualising clusters if numerous patients or codes are fed in, limiting our ability to understand critical high-risk medical codes.

ML researchers have used global feature importance [205] to assess the model's ability to mimic a clinician's diagnostic pathways. Leveraging on the linear relationship between

the embedding space and the context vector, the global feature importance is defined as $S^d = (\Sigma_t^T \alpha_t) w_{out} (W_{emb}[d,:] \cdot (\Sigma_t^T \beta_t))$.

5.5.4.3 Statistical Testing for Model Comparison

Following the training and development of RetainEXT, we performed a rigorous statistical comparison between the newly developed RetainEXT and its predecessor RetainEX. To ensure that observed performance differences were robust and statistically significant, we employed the 5x2 cross-validation combined F-test, a method proposed by Alpaydin [206]. This test was specifically chosen due to its suitability for comparing two models under cross-validation settings, providing a more reliable and robust alternative to traditional paired t-tests.

The 5x2 F-test procedure begins by randomly splitting the dataset into two halves. Each model is then trained and tested five times using different data splits. In each iteration, the models are evaluated on distinct subsets of the data, and their performance metrics are recorded. This method tests the model's performance but also reduces the bias that can arise from using a single training-test split.

The F-statistic, which is the output of this procedure, follows an approximate F-distribution under the null hypothesis that there is no difference in performance between the two models. A critical value of $\alpha = 0.05$ was chosen as the significance threshold. If the resulting p-value is smaller than this threshold, we reject the null hypothesis, indicating that the models exhibit significantly different performance. This approach is particularly effective when comparing models with varying numbers of parameters or degrees of freedom, as is often the case with deep learning models.

In this study, the 5x2 F-test was applied to assess the statistical significance of differences across key performance metrics: Sensitivity, Specificity, AUROC, and F1-Score. These metrics were chosen as they provide a comprehensive view of the model's ability to predict hospitalisation risk, accounting for both the model's y and its ability to classify both positive and negative cases correctly.

The results of the 5x2 F-test, including the calculated F-statistics and corresponding p-values, are reported in Section 5.6, where we present a comparison of model performance. This detailed statistical analysis ensures that the observed improvements in RetainEXT are not due to random chance but are instead reflective of the model's predictive power.

Table 5.1: RetainEXT and baseline performances on predicting risk of hospitalisation due to COVID-19

	Cascading LSTM [207]	1D-CNN + LSTM	RetainEXT (1-Time Diff.)	RetainEXT (3-Time Diff.)				RetainEX [201]	RetainEXT (Extra Emb)	F-Test Best RetainEX vs EXT
Emb. Units			200	200	200	128	128	128, 128	128,128	
LSTM Units			64,64	64,64	64, 64	64,64	128,128	128	128,128	
Dense Units			1	1	50, 1	50,1	50,1	1	50,1	
Sensitivity	55.97%	46.66%	44.83%	58.79%	62.67%	66.03%	56.19%	62.10%	54.10%	0.63
Specificity	88.92%	73.25%	99.23%	98.92%	99.24%	98.77%	99.17%	98.65%	99.32%	2.43
Positive Predictive Value	40.60%	19.11%	52.71%	51.02	61.44%	50.71%	56.44%	46.74%	60.25%	1.39
AUROC	72.00%	59.96%	72.03%	78.86%	80.96%	82.40%	77.68%	80.37%	76.71%	0.65
F1 Score	0.47	0.27	0.48	0.55	0.62	0.57	0.56	0.53	0.57	0.96
Run-time (Min/e-poch)	12	3	10	10	12	12	25	13	32	

Table 5.1 illustrates the performance metrics for RetainEXT, its predecessor, RetainEX and other state-of-the-art temporal classification models. The smaller embedding space with 128 units significantly improves model sensitivity. Additionally, F-test scores for AUROC are below the F-score at the 0.05 critical value, which suggests the null hypothesis can be rejected and we can assume a significant difference between RetainEX and our model, RetainEXT.

5.6 Profiling risk enhancing events

The study cohort comprised 2,277 female and 2,071 male patients with an average age of 69.7 years. Overall, the high-performing risk prediction model was able to identify patients susceptible to hospitalisation due to COVID-19 before being hospitalised and suffering from life-long morbidities. Using ten years of historical GP interactions and a large sample of heterogeneous EHR, our interpretable Temporal Neural Networks (TNN), RetainEXT, shows statistically significant improvements in AUROC, F1 score, sensitivity, and specificity compared to RetainEX and other state-of-the-art TNN. Of note, without sample weighting, the compared models severely over-fit; hence, we trained all models with sample weighting.

5.6.1 Model Performance

Insights from previous Retain iterations helped set a hyperparameter space for optimisation using a hyperband tuner.

The findings show that excluding dropout and L2 regularisation significantly improved model performance. Iterations of RetainEXT with 0 and 0.4 dropout resulted in a 0.17 reduction in F1 score, indicating the data may be very heterogeneous and thus weighting of features

is sparse. The model requires multiple features to assess risk. Similar to the original implementation, convergence on local minima occurred quickly, and dropout required a longer training time to work best. This indicates the possibility that the model is stuck at a local minimum and needs longer to diverge.

In table 5.1, the cascading LSTM [207] has been shown to perform well in imbalanced scenarios, albeit less interpretable. Its sensitivity is comparable to RetainEX and the RetainEXT models, where the only models surpassing it have deeper architectures. The cascading LSTM was trained for 200 epochs compared to 20 epochs for the RetainEXT model, which further emphasises the significance of the attention pathways in the convergence of optimal features.

The 1D-convolutional LSTM leverages both spatial and sequential learning, and its interpretability is facilitated by extracting the kernel weights. However, despite training for over 200 epochs, the F1 score remains relatively low. This suggests that the model's approach of successively aggregating medical codes with a fixed kernel size overlooks the broader trajectory of the patient's condition, assuming that each visit is isolated and failing to account for the potential long-term impact of those medical codes.

Further, adding two extra time representations to the time input layer results in a 5.5% increase in sensitivity, increasing the model's ability to pick up patterns in the interval between GP visits. Including time, features normalised between 0 and 1 generate smaller weight updates on back-propagation, which drives the model closer to the global/local minima.

We also note that RetainEXT outperforms RetainEX in both stacked bi-LSTM configurations, achieving a 0.09 higher F1 score in the optimal configurations. The added LSTM and dense layers allowed the RetainEXT model to understand complex non-linear patterns in the patient's EHR. Both models had similar sensitivity, specificity and AUROC. To determine if the improved performance of RetainEXT significantly differs from other models, we conducted a 5×2 Cross-Fold F-test and calculated an F-score for each evaluation metric. DL models tend to have higher degrees of freedom; hence, a critical value of 0.05 was chosen, and the F-score threshold was set at 1.00.

Sensitivity, AUROC and F1-Score all produced an F-value below the threshold of 1.00, which suggests our implementation of RetainEXT is significantly different from RetainEX. Both Positive Predictive Value and specificity yield an F-score that surpasses the threshold, As the Positive Predictive Value (PPV) and specificity of both models produce F-scores that exceed the threshold, they are within a margin of error in these metrics. This suggests both models have a similar capability to learn low-risk features present before hospitalisation. In

contrast, RetainEXT is more sensitive to high-risk visits or medical codes, allowing for early intervention and possibly reducing the risk of hospitalisation.

The increase in performance between RetainEX and our model is mainly attributed to the stacked LSTM layers and the time-interval representations. RetainEXT provides the added ability to understand the patient's risk after each GP visit, which is critical for the frequency analysis. Our improved model discovered that patients with multiple secondary care referrals are at high risk of being hospitalised due to COVID-19. Additionally, the sample weights have assisted in learning from instances that make up 1.59% of the dataset.

5.6.2 Global Feature Importance

Previous studies have reported that older age and underlying comorbidities, such as hypertension, diabetes and cardiovascular diseases, are risk factors for patients admitted due to COVID-19 [208]. Congruent with their findings, the average patient age in our dataset is 69.8 years, and care home administrative codes were observed in the global feature importance. Elderly patients have poorer immune responses and experience an increased number of age-related comorbidities and therefore are at higher risk of hospitalisation from COVID-19 or other infections [209, 210].

The most important and prevalent category of markers was the hospital service utilisation group, which could include anything from receiving a letter from a specialist to seeing a respiratory physician (see Fig. 5.2). This suggests the importance of administrative codes in understanding the patient's journey through the healthcare system, where the quantity and frequency of referrals may reflect the severity of the condition.

Interestingly, the involvement of a nutritionist or the use of supplements also helped predict hospitalisation risk. This could be due to the correlation between nutrition and immunity [211]. The supplement ferrous fumarate, a prescription for anaemia, was also strongly associated with hospitalisation. It is established that anaemia leads to severe outcomes, which is exacerbated due to COVID-19 [212, 213].

Features in the Physician Diagnostic Thinking (PDT) group are clinical tests to assess the patient's state resulting from a clinician's suspicion. These tests are generally conducted when a physician suspects an infection without knowing the cause.

Consistent with literature [190, 191], we identified white cell differential count, creatinine phosphate and lymphocyte proportion as important markers. A patient usually undergoes a Full Blood Count (FBC) just before admission or during the hospital stay. As a result, these

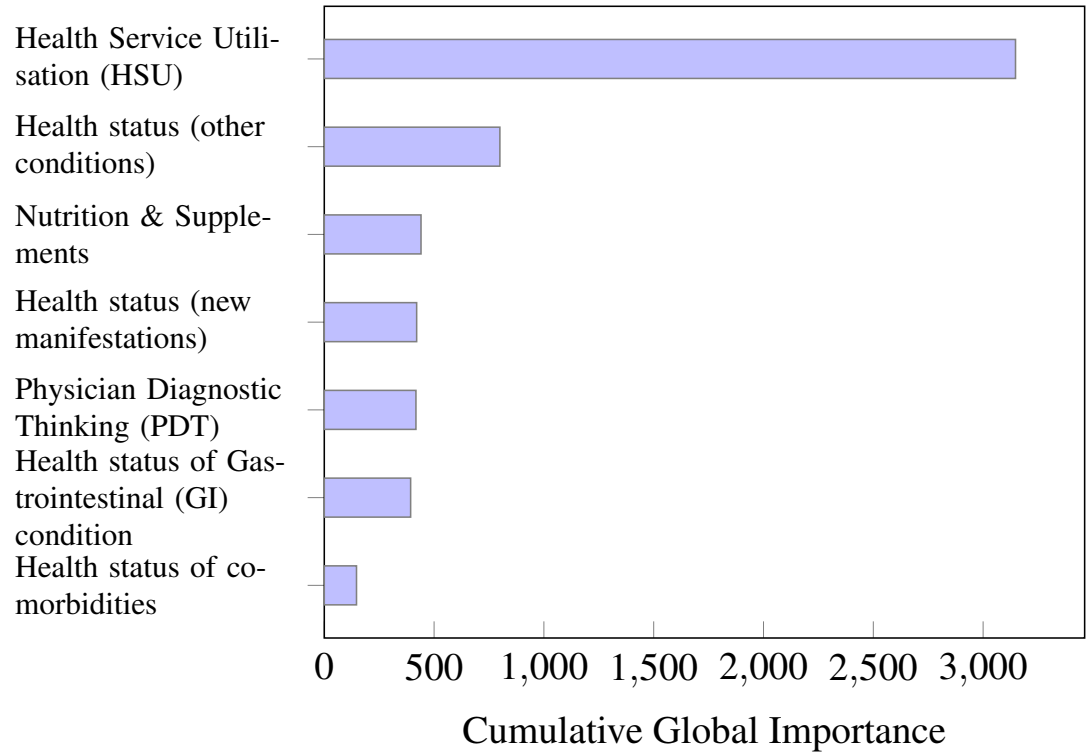


Figure 5.2: Global grouped feature importance for predicting the risk of hospitalisation due to COVID-19. HSU group features are significantly more important as the patient will have multiple interactions with the NHS before being hospitalised. Of note, the high importance of Nutrition & Supplements may relate to the patient's diet playing a vital role in mitigating the risk of adverse outcomes.

markers are strongly associated with hospitalisation.

Finally, the frequency analysis showed that most patients have had telephone interactions with NHS staff or have received a specialist's letter before being hospitalised. This underscores the significance of the Health Service utilisation group markers.

5.6.3 Limitations

- The available data is not well documented, hence, confounding codes may be present in the prediction model and overestimate feature importance.
- The models compared with RetainEX were trained using an encoded binary vector for each visit and may differ in performance and time to convergence.
- The use of global feature importance requires all embedding dimensions to be added to-

gether. This resembles an estimate of the significance of the feature, and further analysis of the latent space would be required for deeper insights into the value of features.

5.7 Summary

From the initial phases of the pandemic, COVID-19-related hospitalisations have heavily overwhelmed the healthcare systems, whose effects are still felt today. With the resurgence of COVID-19 infections and reinfections globally, healthcare resources must be used effectively. Currently, the state-of-the-art model only achieves a 61% accuracy in predicting hospitalisations due to COVID-19.

We conducted fundamental improvements on an interpretable TNN, RetainEX, and provided additional tools to learn from imbalanced EHRs when predicting the risk of adverse outcomes. We leveraged the predictive power of RNN and combined it with a sophisticated attention generation process. This resulted in a better performing model than the current best;RetainEXT achieving 82.40% AUROC on assessing risks of hospitalisation due to COVID-19. The model identified key features, including the importance of telephone calls with patients to assess their severity, poor GI conditions and low levels of ferrous fumarate, indicating anaemia, consistent with existing literature. Additionally, PDT pathways provide valuable insights into possible infections associated with hospitalisation due to COVID-19.

In sum, besides outperforming existing models in predicting COVID-19 hospitalisations, our model was equipped with an additional layer of interpretability, allowing clinicians to use the algorithm for insights on which features were important in the model's predictions.

Future work includes developing visualisation tools to better interpret and apply the findings to more diverse medical records. This increases the model's reliability and helps understand other rare diseases. Additional improvements include using a cascading architecture and custom loss functions that reward early diagnosis. We believe the lessons from this study can guide future researchers in building interpretable recurrent neural network models and contribute to existing literature.

5.8 Connecting the dots and remembering the why

The research presented in this chapter forms a crucial part of our work, contributing to our understanding of how AI and data mining techniques can be applied to address real-world health crises, including the COVID-19 pandemic.

Using self-attention mechanisms to analyse EHRs and predict COVID-19-related hospitalisations represents a significant advancement in the field. Not only does this approach improve the accuracy of predictions, but the interpretability of self-attention mechanisms also provides valuable insights into the factors contributing to these predictions. This can potentially guide clinical decision-making, optimise resource allocation, and improve patient outcomes.

Furthermore, this research shows the importance of collaboration between medical experts and AI scientists. This collaboration ensures that the models we develop are technically sound, clinically relevant, and usable in practice.

Chapter 6

Ontological Osmosis: Infusing Structure into Language Understanding

Contents

6.1 Hidden medical knowledge

6.2 Language models and their clinical utility

Language models, such as the Transformer architecture and its variants, have demonstrated remarkable success in various natural language processing tasks, including text classification, sentiment analysis, and question answering. These models learn and generate human language by training on vast amounts of text data, capturing intricate patterns and relationships within the language. In recent years, the application of language models in the clinical domain has gained significant attention due to their potential to extract valuable insights from unstructured medical data, such as clinical notes, discharge summaries, and patient reports.

With a significant proportion of EHR being recorded as unstructured text, language models are especially relevant and valuable for processing and analysing large volumes of text. Through such models, healthcare professionals can now access and analyse the wealth of information stored within medical text data that was previously resource and time-intensive [214].

Language models have transformed medical text analysis, offering healthcare professionals valuable insights for improved patient care and research. These models support clinical decision-making, patient risk stratification, pharmacovigilance, and clinical trial recruitment [215]. They extract relevant information from patient records, including medical history, prescriptions and test results, to identify at-risk individuals, monitor adverse drug events, and match candidates to clinical trials [216]. Additionally, language models contribute to personalised medicine by integrating diverse data sources to suggest appropriate interventions [217]. This comprehensive approach enhances diagnostic accuracy, treatment efficacy, and research progress, particularly beneficial for rare disease management.

Despite the promising applications of language models in the clinical domain, some challenges need to be addressed, primarily the interpretability of these models [218]. Despite the growing complexity of language models, there has not been much advancement in making the

inner workings of the models more transparent to end users. Therefore, understanding how they arrive at specific predictions or decisions remains difficult, especially for the intended end users of the model. This lack of transparency can hinder trust and adoption among healthcare professionals, who require clear explanations for the recommendations provided by these models [219].

Another challenge is the potential for bias in language models. If the training data contains biases, such as under-representing specific patient populations or historical biases in medical records, the models may perpetuate these biases in their predictions [219]. Hence, ensuring fairness and equity in applying language models in healthcare is crucial to prevent disparities in care and maintain public trust.

Furthermore, integrating language models into clinical workflows requires careful consideration of data privacy and security [218, 219]. Medical data is highly sensitive, and strict regulations govern its use and sharing. Developing language models that can operate within the constraints of data protection laws and ethical guidelines is essential to ensure patient confidentiality and maintain the integrity of the healthcare system [218, 219].

Despite these challenges, language models have unprecedented potential to revolutionise healthcare delivery. By harnessing the power of these models, healthcare professionals can access valuable insights from unstructured medical data, enabling earlier detection of diseases, more accurate diagnoses, and personalised treatment plans. As research advances, it is crucial to address these challenges to realise the benefits of language models in clinical practice.

6.3 Graph Neural Networks and their role in modelling structure

Graph Neural Networks (GNN) have emerged as a powerful tool for modelling complex relationships and dependencies in structured data [220]. Unlike traditional neural networks that operate on fixed-size inputs, GNN can handle graph-structured data, where nodes represent entities and edges represent the relationships between them. This makes GNN particularly well-suited for capturing the intricate patterns and interactions in various domains, including molecular structures, classifying diseases and knowledge graphs [220].

In healthcare, GNN have shown great promise in modelling the complex relationships between medical concepts, such as diseases, symptoms, medications, and patient characteristics [221]. By representing medical knowledge as a graph, GNN can learn the underlying structure and dependencies, enabling more accurate and interpretable predictions. GNN offer significant advancements in healthcare and rare disease management by processing structured

data as graphs. This capability enables more accurate predictions and personalised care, making GNN increasingly valuable in bioinformatics and medical research, particularly for rare disease diagnosis and treatment.

One of the key advantages of GNNs is their ability to capture both the local and global structure of a graph [222]. GNNs can learn the local patterns and dependencies between neighbouring nodes by applying convolutional operations on the graph. Additionally, the iterative nature of GNN architectures allows for the propagation of information across the entire graph, enabling the model to capture long-range dependencies and global patterns [222].

Another critical aspect of GNNs is their ability to handle a variety of data types [223]. Healthcare data often comes from diverse sources and in various formats, including structured health records, unstructured clinical notes, and imaging data. GNNs can integrate these different data types by representing them as nodes and edges in a unified graph structure. This allows for the joint modelling of multiple modalities and the discovery of cross-modal relationships [224].

However, as evidenced, the application of GNNs in healthcare also faces challenges, a key problem of being not interpretable by non-technical users [224]. As research in this field advances, it is crucial to address the challenges of interpretability to fully realise the benefits of GNN in healthcare applications.

6.4 Augmenting rarity with ontologies

Rare diseases pose a significant challenge in healthcare due to their low prevalence, complex manifestations, and the scarcity of available data [225]. Patients with rare diseases often face delayed diagnoses, misdiagnoses, and limited treatment options, leading to poor health outcomes and reduced quality of life. This is primarily due to the lack of comprehensive data, knowledge and understanding of rare diseases.

Medical ontologies [226], which are formal, structured representations of medical concepts and their relationships, can play a crucial role in augmenting the understanding and management of rare diseases. These structured representations of medical knowledge enable improved disease characterisation, facilitating the identification of similarities and differences between rare conditions, and can help bridge the gap between the available information and the clinical decision-making process. Ontologies also promote data sharing and collaboration among researchers and clinicians by providing a common language for describing rare disease concepts.

A significant challenge is the integration of ontologies into clinical workflows and decision support systems. Seamless integration of ontologies into the tools and platforms used by healthcare professionals can provide them with easy access to relevant knowledge and guidance at the point of care. This requires the development of user-friendly interfaces and the integration of ontologies with existing EHR systems and clinical decision support tools.

Furthermore, adopting ontologies in managing rare diseases requires a shift in the mindset and practices of healthcare professionals. Training and education initiatives are necessary to raise awareness about the value of ontologies and to equip clinicians with the skills and knowledge needed to effectively utilise these resources in their daily practice.

Despite these challenges, the potential of ontologies to revolutionise the management of rare diseases is significant. By augmenting the understanding and knowledge of these conditions, ontologies can contribute to earlier and more accurate diagnoses, developing targeted therapies, and improving patient outcomes. As research advances, it is crucial to foster collaboration between ontology developers, researchers, and clinicians to fully realise the benefits of ontologies in understanding and diagnosing rare diseases.

6.5 The need for structure in language

The COVID-19 pandemic has exacerbated and exposed challenges and limitations in global healthcare systems, such as the rapid stratification of patient risk, data interoperability for real-time insights, and swift adaptation to surges in patient volume [227]. There remains an urgent need for proactive approaches in preventative medicine, particularly the development of interpretable risk stratification tools for early identification of high-risk individuals in primary care settings [94, 166, 228, 229]. Such tools facilitate early intervention, which holds the potential to significantly reduce premature mortality and alleviate the strain on healthcare systems by reducing the demand for secondary care.

ML models have evolved to become diagnostic tools for identifying high-risk patients and predicting outcomes, such as for COVID-19 [230]. However, such models were not known to be capable of capturing the complex interdependencies between medical conditions and the patients' evolving medical status. Nonetheless, part of the challenge comes from the fragmented data landscape resulting from non-standardised coding systems nationally and globally, especially in primary care settings [231, 232].

To address these challenges, recent advancements combined Natural Language Processing (NLP) with ML using unstructured clinical texts in EHR to further collaborative under-

standing of the conditions [233]. However, complex black-box predictions raise concerns around reliability and fairness in high-stakes medical decision-making. Including structured ontological knowledge to regularise learning has shown promise for improving predictive accuracy and enabling the interpretation of model behaviour in relation to formal relationships. NLP in ML adds layers of reasoning that look beyond the codes used and into more precise, human-understandable descriptions in its analysis and output. Furthermore, GNN have expanded the capabilities of data analysis in healthcare by modelling complex relationships and dependencies among clinical concepts within EHR, which allows the creation of concept-specific graphs [234]. These graphs learn from the hierarchical nature of medical codes and capture evolving patient conditions over time, offering a more profound understanding that traditional models lack [235].

Building upon these advancements, this chapter introduces a novel architecture that synergises GNN with a pre-trained language model, TinyBERT, to maximise the utility of EHR by exploiting GNN to induce structured embeddings of medical concepts. This enriches TinyBERT's general analysis capabilities with specialised medical domain knowledge.

6.6 The domain of structure in medical language

Decision trees [190–192] and logistic regression [196, 197] offer interpretability for clinical prediction tasks but are limited in using the breadth of information recorded in EHR [196, 198]. Most techniques rely solely on structured inputs like demographics, medications, and diagnosis codes for training predictive models [236] while ignoring valuable signals from unstructured free text in written notes and clinical code descriptions. The deep EHR method addressed this [237] by effectively using both structured and unstructured data for EHR analysis to comprehensively analyse patient data.

Recent ML developments seek to capture complex hierarchical relationships between medical concepts. Choi implemented Directed Acyclic Graph (DAG) to model such relationships explicitly using a static approach [238]. However, DAG poorly represent the nuanced correlation between related medical concepts within the hierarchy and offers limited interpretability into the importance of ancestral codes. Graph convolutional networks (GCNs) take a more flexible approach through neighbourhood aggregation but still assume consistent local transitions.

More advanced Graph Attention Networkss (GATs) allow adaptive, non-linear weighting of edges based on learned relevance between concepts, offering greater flexibility in mod-

elling the relationships between nodes. This provides greater representational power to model the precise importance of specific diseases to their ancestors. However, unconstrained attention risks losing generalisability. In contrast, Graph Convolutional Networks (GCN) enforce an equality-based smoothing that aids in higher-level abstractions along broad taxonomy dimensions. Therefore, finding the right balance between precisely modelling local ontological nuances with GAT and smoothing out their global categorisation with GCN is crucial for advancing clinical modelling [239].

Since then, NLP has been incorporated to leverage unstructured clinical text and improve model generalisability. Knowledge graphs generated using GNN map connections between medical codes based on expert-defined ontologies have improved model generalisation and robustness compared to traditional feature-based methods [240]. Combining knowledge graphs with large pre-trained language models like BERT [241] has further improved by enabling models to ingest and analyse unstructured data. Nonetheless, they have scarcely been adopted, as large transformer models are computationally intensive. An alternative is to use less computationally intensive models like TinyBERT [242], though this lacks inherent clinical knowledge capabilities.

However, despite these advancements, current approaches still have limitations. Recent literature highlights that these techniques have focused narrowly on medication or procedure prediction rather than risk assessment and screening applications [243, 244]. Furthermore, the graph constructions often rely on predefined taxonomies rather than data-driven learning of flexible concept correlations [245, 246]. This highlights the need for more comprehensive approaches to address risk assessment and screening while incorporating data-driven learning of concept correlations.

This study proposes a novel inductive graph representation learning technique to address these limitations. By combining GAT and GCNs, we aim to induce medical concept embeddings that encode ontological relationships within TinyBERT. This approach integrates structured knowledge graphs with an efficient pre-trained model, potentially overcoming the computational limitations of larger transformer models while retaining clinical knowledge capabilities. Furthermore, adhering to the Bidirectional Encoder Electronic Health Records Representations from Transformers (BEHRT) training scheme [247] enables the joint use of primary care Read codes and their definitions for predictive tasks, potentially facilitating early diagnosis and interventions. This innovative approach seeks to bridge the gap between computational efficiency and comprehensive risk assessment in clinical applications.

6.7 Modelling medical hierarchies in deep learning

This study uses a retrospective longitudinal case-control design following patient journeys in primary care data from the Secure Anonymised Information Linkage (SAIL) databank [248]. SAIL contains approximately 80% coverage of the Welsh population's primary care records.

6.7.1 Cohort creation

Using the primary care data in SAIL, 3 datasets were created: COVID-19 hospitalisation, COVID-19 mortality, and stroke (Table 6.2). Patients were included if they were aged between 18 and 100 at the start of the study period (01/06/2017) and had at least two GP interactions during the study period (01/06/2017 to the event date). Features extracted included diagnosis, procedural and medication codes, demographics, and inter-visit times.

Patients who died within 15 days of hospitalisation were excluded from the COVID-19 hospitalisation and added into the COVID-19 mortality dataset [180]. Patients who had stroke-related diagnoses within the study period were included in the stroke dataset. These datasets provide diverse aspects and scales of healthcare data for robust model evaluation.

The event date is unique for each dataset. It refers to the hospitalisation date that is followed by a positive antibody or Polymerase Chain Reaction (PCR) test for the COVID-19 hospitalisation dataset; the mortality date after a hospitalisation accompanied by a positive antibody or PCR test for the COVID-19 mortality dataset; and, the earliest date of stroke-related diagnosis within the study period for the stroke dataset.

For each dataset, the read codes were filtered to have a minimum of 5 occurrences, reducing noise. Longitudinal primary care data up to the event date was used, along with Read codes, demographics like age and gender, and the number of days between visits as features. The data was split 80/20 into training and hold-out testing sets, respectively. Cases and controls were determined by read codes (available upon request).

Stroke Dataset and Configuration The stroke dataset used in this study is specifically designed to evaluate the model's performance on a real-world clinical scenario. For this dataset, we explored configurations with pre-trained embeddings from TinyBERT, using text-based and graph-based representations of the medical data. Table 6.1 provides a detailed breakdown of these configurations, showing the interplay between pre-trained embeddings, textual information, and graph representations in improving classification performance for stroke-related outcomes.

6.7.2 Stratified K-Fold

Stratified k-fold cross-validation is employed by dividing the training data into 10 class-balanced subsets, each used once as a validation fold while training on the other 9 aggregated folds. This controls for sampling bias and provides more reliable generalisation estimates than conventional splits. Evaluation metrics are computed on the validation folds and an initially held-out test set, then averaged across folds to reduce variance. By assessing many combinations of data subsets, k-fold cross-validation provides a regularised evaluation of model robustness while final testing on new data measures real-world applicability. The multi-fold averaging provides reliability superior to single train-test passes [249].

6.7.3 Modelling

An extensive ablation study evaluated multiple TinyBERT model configurations on the three datasets. Each configuration selectively incorporated the following parameters: 1) Pre-trained vs newly initialised word embeddings for general prior knowledge, 2) additional graph embeddings encoding medical ontology relationships for domain-specific clinical insights and 3) diagnosis/medication code embeddings vs textual definition embeddings for interpretability. The ablation study shows the individual and combined effects of the parameters on the model's predictive performance and determines the best combination to maximise model performance.

6.7.4 Model Architecture

The study uses a novel architecture which combines a language model, TinyBERT, with a sophisticated GCN and Graph Attention v2 Convolution (GATv2Conv). This architecture is designed to manage the complexity of healthcare data, including unstructured clinical text and structured medical entities such as diagnoses and medication codes.

6.7.4.1 Hierarchical Embeddings

The diagnosis and medication ontology graphs have nodes, V_{diag} and V_{med}, representing unique medical conditions and medicine codes, and edges, E_{diag} and E_{med}, capturing hierarchical relationships between the codes based on established taxonomies.

Parent-child connections directly linking medical conditions are modelled using GATv2Conv, which allows flexible connectivity patterns. This enables selectively highlight-

ing the most pertinent local relationships for targeted encoding into diagnosis embeddings E_{diag} and medicine embeddings E_{med}.

However, when modelling multi-hop ancestor relationships, GCN are leveraged, which smooth over lineages through uniform neighbour aggregation schemes to regularise the abstraction of medical concepts in E_{diag} and E_{med}.

By combining these complementary strategies, we simultaneously distil precise semantic connections and higher-order taxonomic knowledge into the diagnosis embeddings $E_{diag} \in \mathbb{R}^{|V_{diag}| \times D}$, injecting domain structure into the learned representations. This can be formulated as:

$$E_{diag} = \sigma(\text{GATv2Conv}(V_{diag}, E_{diag}) + \text{GCNConv}(V_{diag}, E_{diag})) \tag{6.1}$$

Where σ is an activation function and D is the embedding dimensionality.

6.7.5 Journey of Patient Data through the Algorithm

Tokenization and Data Preparation Drawing inspiration from the BEHRT model [247], each patient visit is tokenised into a sequence including clinical text, diagnosis, and medication codes. The sequence is prefixed with a [CLS] token and suffixed with a [SEP] token. Mathematically, a visit V with N tokens is:

$$V = [CLS, t_1, t_2, \ldots, t_N, SEP] \tag{6.2}$$

Multitude of Embeddings The tokenized sequence is enriched by a variety of embeddings:

- Word Embeddings, E_w
- Token-Type Embeddings, E_t
- Diagnosis Embedding, E_{diag}
- Age Embeddings, E_a
- Medicine Embedding, E_{med}
- Gender Embeddings, E_g

Embedding Aggregation Several techniques for embedding aggregation were explored, including gated fusion, bilinear pooling, concatenation, and attention fusion. The final token-level representation generated was:

$$E_{\text{token}} = \text{AggregationFunction}(E_w, E_{\text{diag}}, E_{\text{med}}, E_t, E_a, E_g) \tag{6.3}$$

Processing through TinyBERT The token-level embeddings E_{token} were processed through the TinyBERT model, a lightweight yet powerful language model. The model applied self-attention mechanisms, yielding the output O:

$$O = \text{Self-Attention}(E_{\text{token}}) \tag{6.4}$$

These outputs were aggregated to produce the patient-level representation E_{patient}:

$$E_{\text{patient}} = \sum_{i=1}^{n} O_i \tag{6.5}$$

6.7.6 Fine-tuning for the Classification Tasks

An advantageous decision shown by the ablation study was to use the pre-trained embeddings from TinyBERT in the model to provide a rich representation of each GP visit. This is vital when working with sequential multi-visit data, especially in classification tasks.

Leveraging TinyBERT's Pre-trained Embeddings The model capitalises on the existing pre-trained Transformer encoder, TinyBERT, removing the need to manually pre-train the model. This encoder adeptly captures the nuances of individual medical visits, producing the visit-level embedding, v_τ^*, at time τ.

From Visit-Level Representation to Classification Using the visit-level embeddings, v_τ^* derived from TinyBERT, we construct a comprehensive representation for the entire patient history, symbolised as E_{patient}. This representation is then processed through a fully-connected layer with dropout, and a class label is derived using the softmax function:

$$\text{Class Label} = \text{Softmax}(\text{Fully-Connected}(E_{\text{patient}})) \tag{6.6}$$

Technical Contributions The contributions of this work include the novel integration of embedding aggregation layers and the use of metrics based on distance to evaluate the proximity of augmented infrequently coded terms. This novel approach allows better model interpretability and relevance to clinical decision-making. Additionally, combining TinyBERT, GCN, and GATv2Conv enhances the model's ability to process unstructured and structured healthcare data, bridging the gap between large language models and clinical knowledge.

6.7.7 Interpretability

Multiple complementary approaches provide transparency into the predictive modelling of patient trajectories.

Class Activation Maps (CAM) visually highlights influential terms in the sequence-based transformer predictions. The highlighted medication usage and risk factors align with expected clinical indicators for the target conditions. This supports qualitative accountability regarding the model's data-driven rationales.

Additionally, graph edge attention analysis quantifies the relevance of medical code relationships, indicating a predominance of local parental dependencies in the model's reasoning. The attention flows exhibit limited global propagation across broader ontology categories. Self-loops indicate reliance on specific individual codes independent of their ancestors or descendants.

Lastly, subgroup evaluation of metrics across age bands and gender categories helps audit for potential unwanted correlations outside the core hierarchy encoding. Monitoring for emerging biases will remain critical as clinical understanding evolves amidst new evidence on risk factors.

Through combining sequence relevance tracing, modular graph accountability, and algorithmic fairness analyses, these interpretation methods improve trust in model rationales based on causal mechanisms rather than spurious correlations. The layered insights enhance transparency of the model's decision-making process.

6.7.8 Implementation

The model was trained using the PyTorch DL framework. The GATv2Conv was implemented using the PyTorch Geometric library using a 2-layer GATv2Conv with hidden dimensions set to 32. The TinyBERT model was initialised with pre-trained weights from the HuggingFace library [250]. Adam optimiser was used with a learning rate of 0.001 and weight decay of 5e-4. Finally, each model variation was fine-tuned on each dataset for 10 epochs with a learning rate of 5e-5 and a Cosine scheduler with a batch size of 8.

6.8 Results and Evaluation

A total of 977, 18,766, and 47,132 patients were included in the COVID-19 Hospitalisation, COVID-19 Mortality, and Stroke datasets, as shown in Table 6.2.

Table 6.1: Best Performing Models on Different Datasets

	CV19 Hospitalisation		**CV Death**		**Stroke**	
Pre-trained	Yes	Yes	Yes	Yes	Yes	Yes
Text	No	Yes	No	Yes	No	Yes
Graph	Yes	No	Yes	No	Yes	Yes
Loss	0.76 ± 0.02	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.25 \pm 0.01}$	0.27 ± 0.01	0.60 ± 0.02	$\mathbf{0.59 \pm 0.01}$
Accuracy	0.52 ± 0.00	0.52 ± 0.00	0.90 ± 0.01	0.90 ± 0.01	0.71 ± 0.00	0.71 ± 0.00
Precision	0.53 ± 0.00	0.53 ± 0.00	0.89 ± 0.01	0.89 ± 0.01	0.73 ± 0.00	0.73 ± 0.00
Recall	$\mathbf{0.62 \pm 0.03}$	0.59 ± 0.03	$\mathbf{0.89 \pm 0.01}$	0.87 ± 0.02	$\mathbf{0.72 \pm 0.01}$	0.71 ± 0.01
F1	$\mathbf{0.56 \pm 0.01}$	0.55 ± 0.01	$\mathbf{0.89 \pm 0.01}$	0.88 ± 0.01	0.72 ± 0.00	0.72 ± 0.00
Runtime	6.81 ± 0.14	4.50 ± 0.12	0.29 ± 0.01	0.40 ± 0.00	17.98 ± 0.07	17.80 ± 0.07

Table 6.2: Summary of Datasets

	CV19 Mortality		CV19 Hosp.		Stroke	
	Control	Case	Control	Case	Control	Case
Male	23.95%	18.12%	18.74%	19.28%	27.05%	28.57%
Female	30.81%	27.12%	30.53%	31.45%	21.65%	22.73%
Avg Age	79.93	82.12	48.26	48.30	70.97	71.51
Total	977		18,764		47,132	

The impact of pre-trained embeddings, text, and graph structures on three different medical datasets (COVID-19 Hospitalisation, COVID-19 Mortality and Stroke) was evaluated during the assessment of model performance with various configurations.

The summarised results of the best-performing models on these datasets are shown in Table 6.1. Overall, the models employing graph structures consistently achieve a lower loss than their counterparts using text. This demonstrates the significance of graph embeddings in improving classification capabilities. The accuracy remained consistent across datasets irrespective of the use of text or graph structures. Notably, the highest accuracy achieved (approximately 0.90) was in predicting COVID-19 mortality. The precision and recall figures were relatively similar across the three datasets, with the F1 score reflecting these values, suggesting balanced false positives and false negatives. In terms of runtime, predicting for stroke was more computationally intensive (estimated 18 seconds), irrespective of using text or graph structures.

Incorporating graph structures for the COVID-19 hospitalisation dataset improved the loss metric from 0.76 to 0.72 while other metrics remained consistent. This suggests that while the model's overall accuracy and predictive capabilities remained the same, the certainty of its predictions improved with the inclusion of graph structures. Models employing graph struc-

tures for the COVID-19 mortality dataset demonstrated a marginal improvement in loss and recall, while other metrics remained consistent. The high accuracy of approximately 0.90 for both model configurations highlights the robustness of the models in predicting COVID-19 Death. Similar to COVID-19 hospitalisation, the use of graph structures for the stroke dataset improved the model's loss from 0.60 to 0.59. Other metrics remained consistent, suggesting a minor enhancement in the model's prediction certainty.

6.9 Discussion

This study offers novel insights into classifying patients with various health outcomes using a transformer-based model fine-tuned on patient sequences encompassing clinical text, diagnosis, and medication codes.

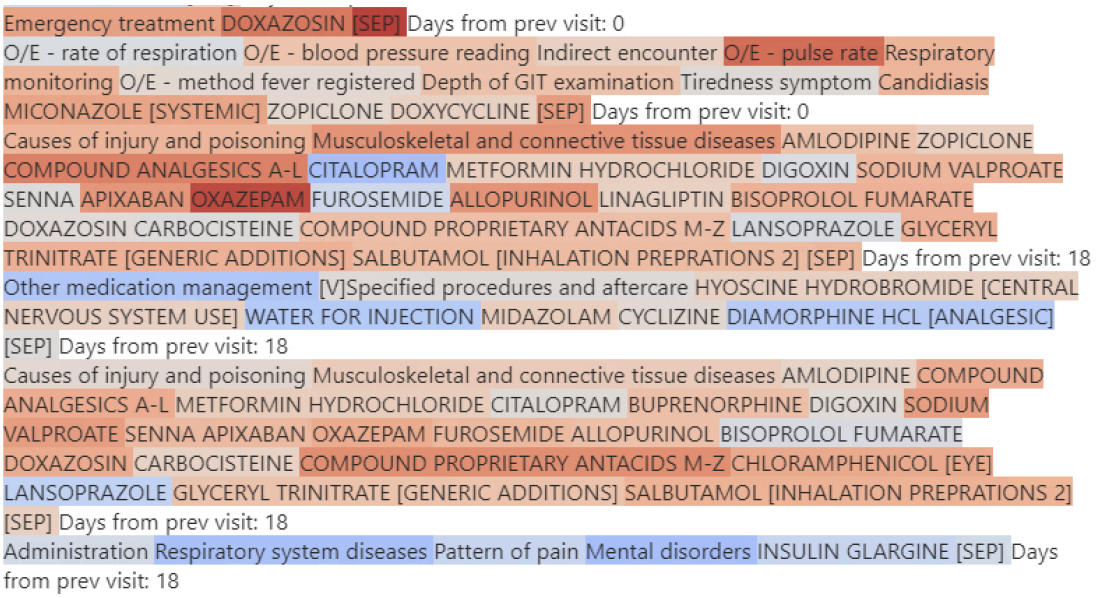


Figure 6.1: CAM Salience over the words of each visit. Red denotes a feature that increases the risk of mortality for patients with COVID-19

The model demonstrated varying performances when classifying patients with COVID-19 hospitalisation, COVID-19 mortality, and stroke. The fine-tuned transformer showed improvements with accuracy reaching up to 0.90 ± 0.01 which suggests the model's ability to understand complex patient data. It was evident that including text and graph-based data improved the model's understanding of the context, resulting in improved performance.

Using CAM correctly associated medications and clinical indicators aligned with the health outcomes. For instance, Bisoprolol Fumarate, often used to manage hypertension, suggests possible cardiovascular issues in the stroke dataset [251]. While a medication does not confirm a diagnosis, it may be an early indicator of a diagnosis, requiring and enabling early clinical interventions.

Additionally, using [SEP] at the end of each visit allows the model to perceive and display the aggregated risk of that visit, considering all the conditions and medication previously analysed. Nonetheless, model interpretations should be supplemented with thorough clinical evaluations. The fine-tuning process was instrumental in adapting the transformer model to the specific nature and requirements of the data. By training on patient-specific visit sequences and using token-level embeddings, the model captured intricate details and temporal patterns inherent in the patient journey, highlighting the importance of tailoring pre-trained models to domain-specific context.

6.10 Summary

This study presents a robust approach to predicting health outcomes using patient visit sequences. Through the integration of fine-tuning, multi-level embeddings, and careful interpretation methods, a framework that can be adopted and adapted for various clinical prediction tasks has been laid. While the journey through the maze of healthcare data is complex, tools like the one developed illuminate possible paths, thus improving patient care and outcomes.

By leveraging on detailed patient trajectories, healthcare professionals can receive early warnings or predictive insights into a patient's potential health risks. Further, the model highlighted specific tokens that could aid clinicians in making timely and better-informed patient interventions.

6.10.1 Limitations and Future Work

While the model performed well across the datasets, it is essential to consider potential biases in data collection and representation. The datasets used in this study are derived from primary care records, which may have limitations in terms of completeness, representativeness, and the under-representation of certain populations, such as those with rare diseases or non-English speaking backgrounds. Additionally, the data primarily consist of structured codes and clinical text, which may lack some of the nuanced medical details that could influence predictions.

External validation on diverse datasets, including data from different healthcare systems or patient demographics, would help establish the model's robustness across various clinical contexts. Future work could also integrate more sources of patient data, such as imaging or genomics, which could offer a richer understanding of patients' medical histories and help improve model performance by capturing complex biological factors that may not be fully reflected in the clinical data alone.

6.11 Connecting the dots and remembering the why

This research aims to contribute to this ongoing transformation by harnessing the power of data-driven approaches to detect and manage rare events in healthcare, while ensuring that the methods used are clinically sound, interpretable, and actionable for healthcare professionals.

The importance of this research lies in demonstrating the potential of combining the strengths of natural language processing, graph neural networks, and medical ontologies. This study sought to develop a framework that effectively utilises the vast data available in EHRs to enable earlier diagnosis and intervention.

By capturing the heterogeneity and variability of rare diseases at the individual patient level, the proposed framework can support the tailoring of treatment strategies to each patient's specific needs and characteristics, which could improve patient outcomes. Finally, this study considered ways to allow the output of the models to be more interpretable for clinicians, which has been a challenge for ML models in the medical domain.

Reflecting on the clinical journey, it is important to acknowledge the invaluable work of healthcare professionals in managing complex medical conditions. This study highlights the importance of structured and interpretable tools that assist clinicians in making data-driven decisions. Integrating advanced models like the one presented here can provide valuable insights without replacing clinical judgement but complementing clinicians' expertise. By situating this work within the broader context of clinical practice, we aim to enhance decision-making processes, particularly in rare diseases, by offering predictive insights and the interpretability necessary for healthcare providers to trust and act upon the results.

In sum, this study met its objective of introducing a novel architecture that synergises GNN with a pre-trained language model, TinyBERT, to maximise the utility of EHR by exploiting GNN to induce structured embeddings of medical concepts while developing a way to improve interpretability for clinical users.

Findings of this study show the potential of using the said architecture in improving rare disease diagnosis while demonstrating the feasibility of making complex models transparent and interpretable for clinicians.

Chapter 7

Decoding Rarity: Machine Learning to Distinguish Complex Cardiac Diseases with Similar Presentations

Contents

7.1 Addressing Ambiguities in Diagnosis

Diagnosing rare diseases is challenging, especially when their symptoms overlap with more common conditions. Symptoms such as fatigue, chest pain, and shortness of breath are often attributed to common conditions like coronary artery disease or HCM. However, these symptoms may also indicate FD, a rare condition, often misdiagnosed or diagnosed too late due to its ambiguous presentation. This phenomenon, known as ambiguity imbalance, occurs when rare diseases are overlooked or misdiagnosed because their symptoms mimic those of more common diseases, leading to diagnostic delays and inappropriate treatments that worsen patient outcomes.

Ambiguity imbalance refers to the diagnostic challenge of distinguishing rare diseases from more common conditions, especially when symptoms overlap. This imbalance can lead to

a higher rate of misdiagnosis for rare diseases, resulting in delayed treatment and worsened outcomes. In machine learning, addressing this imbalance involves improving the model's ability to accurately classify rare diseases amidst a preponderance of common conditions.

Research shows that patients with rare diseases often face years of diagnostic delays, being passed between specialists and undergoing multiple tests before a correct diagnosis is reached [12]. These delays not only prolong suffering but can result in misdiagnoses and harmful treatments. Addressing the ambiguity imbalance with better diagnostic tools, such as machine learning (ML), is crucial for improving diagnostic accuracy and patient outcomes.

Our research transitions from addressing class imbalances in datasets to tackling ambiguity imbalance in rare disease diagnosis. FD, in particular, exemplifies the difficulty of distinguishing it from other diseases, particularly HCM, which shares many symptoms. This chapter discusses how ML can mitigate ambiguity imbalance by using routine cardiac tests and clinical data to more accurately differentiate between FD and HCM, providing faster and more reliable diagnoses.

7.1.1 A Zebra in a Herd of Horses

In medical training, 'zebra' denotes a rare disease, unlike the more common horses." The saying "When you hear hoof beats, think of horses, not zebras" encourages clinicians to prioritise common diagnoses when faced with typical symptoms [252]. While this approach works well for most cases, it can be problematic when a rare disease, such as FD, presents symptoms that overlap with more common diseases like HCM.

Focusing on common conditions can lead to misdiagnosis and delayed treatment for rare diseases, the "zebras". Many FD patients are initially diagnosed with HCM due to overlapping symptoms like left ventricular hypertrophy and arrhythmias. The real-world challenge is that clinicians may overlook rare diseases like FD due to limited data, lack of awareness, and greater familiarity with common conditions. The diagnostic process for FD requires special attention to symptoms that could easily be misattributed to more common diseases. Machine learning can help address this diagnostic imbalance by identifying "zebras" like FD earlier, enabling clinicians to detect these rare conditions sooner.

7.1.2 Fabry Disease: A Rare and Overlooked Condition

FD is a rare genetic disorder caused by mutations in the α-galactosidase A gene (GLA) gene, leading to deficient alpha-galactosidase A activity. It affects approximately 1 in 15,000 in-

dividuals worldwide [253] and presents with a range of symptoms, including chronic pain, fatigue, clouded vision, and cardiac complications such as arrhythmias and Left Ventricular Hypertrophy (LVH). Despite these distinct symptoms, FD is often overlooked due to its rarity and heterogeneous presentation.

A key diagnostic challenge with FD is its heterogeneity. Some patients exhibit classic symptoms early in life, while others experience more subtle or intermittent symptoms that are easily misattributed to other conditions. This variability complicates the standardisation of diagnostic approaches, making it harder for clinicians to promptly identify FD. Moreover, female patients may present with normal enzyme levels, making genetic testing essential for a definitive diagnosis.

Detection of FD is further complicated by sporadic encounters in clinical practice and the similarity of its symptoms to other conditions [254, 255]. As a result, clinicians are more likely to test for common diseases, leading to misdiagnosis or delayed treatment. The diagnosis of FD can take anywhere from 3 to 15 years [256, 257], which can lead to significant health deterioration and increased treatment costs [258]. Early diagnosis is critical to prevent irreversible organ damage. The scarcity of data on FD also complicates the development of diagnostic algorithms, making it difficult to apply machine learning models to a broader population. ML offers a promising solution by leveraging routine clinical data to differentiate FD from diseases with similar presentations.

7.1.3 Hypertrophic Cardiomyopathy: A Common Misdiagnosis for Fabry Disease

HCM is a common genetic heart condition, affecting approximately 1 in 500 individuals. Like FD, HCM is characterised by left ventricular hypertrophy and arrhythmias, making it a frequent misdiagnosis for FD. The challenge lies in the fact that HCM is better known to clinicians, who are more likely to consider it first when patients present with symptoms like chest pain, shortness of breath, or arrhythmias.

In contrast, FD lacks specific diagnostic markers and is often overlooked when symptoms overlap with those of HCM. This results in misdiagnosis and delayed treatment, leading to worse outcomes for FD patients. The misalignment between FD and HCM highlights a fundamental challenge in rare disease diagnosis: clinicians often prioritise more common diseases due to well-established diagnostic tools. Machine learning can help address this challenge by using routine clinical tests to differentiate between FD and HCM even when symptoms overlap

significantly.

7.1.4 The Diagnostic Journey for Rare Disease Patients

The diagnostic journey for many patients begins with primary care visits, where common symptoms like fatigue and chronic pain are often attributed to stress or anxiety. As a result, diagnostic delays are common at this stage, as clinicians may not recognise these symptoms as signs of a rare condition like FD. Patients are referred to cardiologists when cardiac symptoms arise, but FD is still often overlooked due to its similarity with more familiar conditions like HCM.

Only after extensive testing, including genetic testing, is FD typically diagnosed. This diagnostic process, which can take years, not only causes frustration for patients but also increases the risk of irreversible organ damage. The diagnostic journey for FD underscores systemic challenges in rare disease diagnosis, such as fragmented data and a lack of awareness among clinicians. Patients are often passed between specialists, prolonging the time it takes to receive a correct diagnosis. These challenges highlight the need for more efficient diagnostic tools to identify rare diseases like FD earlier.

7.2 Data Collection and Standardisation Challenges

Data collection for FD diagnosis is complicated by fragmented healthcare systems and outdated legacy technologies. For example, ECHO reports are often stored as PDFs in various formats, making extracting useful data for analysis difficult. Moreover, the software required to visualise these reports was limited to a single terminal and required a significant fee to convert the data into usable formats.

Fragmented healthcare systems often result in data that is difficult to standardise, hindering machine learning models' ability to train effectively. For example, ECHO reports stored in non-uniform formats make it challenging to create a consistent dataset. Without standardised data, ML models risk developing biases or missing key patterns that could be crucial for accurate diagnosis. Addressing this fragmentation through custom data extraction tools is essential to ensure that the data fed into ML algorithms is reliable and representative.

The lack of interoperability between healthcare systems and restrictions on data access create substantial barriers to efficient data collection and analysis. Many healthcare systems still rely on outdated technology, making it difficult to gather meaningful clinical data promptly.

To address these challenges, we developed custom tools to automate data extraction and standardise data from fragmented sources. By ensuring data consistency, we aim to facilitate the integration of clinical data into machine learning models, improving the diagnosis of rare diseases like FD.

7.2.1 Control Case Matching

The challenge of diagnosing FD in the real world is reflected in the misdiagnosis of FD as HCM. To address this, we selected HCM patients as the control group for this study, creating a dataset that mirrors the challenges clinicians face in distinguishing between these two diseases in real-world settings. The clinical teams at the University Hospital of Wales played a crucial role in identifying FD and HCM patients based on routine clinical criteria, and data were collected from three routine cardiac tests—echocardiograms, Holter monitors, and ECGs.

Using real-world data, we aim to improve our understanding of the limitations of traditional diagnostic methods and demonstrate the utility of machine learning in diagnosing FD. The goal is to improve diagnostic accuracy and provide clinicians with a reliable tool to differentiate between FD and HCM in clinical practice.

7.2.2 Differentiating FD from HCM

There is a notable gap in evidence regarding FD, particularly in applying ML for its diagnosis. Previous studies have explored ML for FD diagnosis, including the use of CNNs to identify protein anomalies in urine samples [259], medical records [260, 261], and distinguishing FD from HCM using ECG markers [262]. These studies reported classifier performances up to 0.87 AUC, demonstrating the potential of ML in FD diagnosis.

HCM, a common genetic heart condition, is frequently misdiagnosed as FD due to shared symptoms such as left ventricular hypertrophy and arrhythmias [263, 264]. This emphasises the clinical challenge of differentiating between the two diseases. However, the small sample sizes and limited control groups in existing studies, often consisting of healthy individuals or generic medical records [259, 260], have limited the generalizability of the findings.

Further complicating this issue, most studies rely on controls that differ significantly from FD patients, resulting in an oversimplified differentiation. To address this gap, our study uses HCM patients as the control group, better reflecting real-world diagnostic challenges where FD is often confused with HCM. By leveraging routine clinical tests—echocardiograms, Holter

monitors, and ECGs, we aim to improve the differentiation between FD and HCM with ML tools that utilise a more extensive and realistic dataset.

In addition, while ML has proven effective in rare disease diagnosis, its application is hindered by the "black box" nature of many models, which clinicians often struggle to trust [265, 266]. To address this, our study focuses on increasing model transparency and explainability. This will allow clinicians to understand better and validate MLs decisions, ensuring that the model's pattern is clinically sound and fostering trust in its use for diagnosing complex conditions like FD and HCM.

The ultimate goal of this study is to provide a reliable ML tool that improves diagnostic accuracy for FD and HCM by analysing routine cardiac data. In doing so, we aim to fill the evidence gap, enhance transparency, and assist clinicians in making more informed decisions in clinical practice.

7.3 Bridging the Gap Between Rare and Common Diseases in Diagnosis

7.3.1 Study Design and Recruitment

Existing research on FD using ML typically uses control groups consisting of either healthy individuals [259] or generic medical records [260, 261]. However, these groups differ significantly from FD patients, artificially inflating classification accuracy. In clinical practice, FD is frequently misdiagnosed as HCM due to overlapping symptoms [267–269]. Thus, this study selected HCM patients as the control group, reflecting real-world diagnostic challenges.

The clinical teams at the University Hospital of Wales identified participants. Inclusion criteria were a diagnosis of FD or HCM and being over 18 years old. Exclusion criteria included acute illness or post-operative episodes during data collection. Cases of FD were confirmed through genetic testing, while expert cardiologists verified HCM diagnosis based on standard diagnostic criteria [270].

Data were collected from three routine cardiac tests: echocardiograms (echo), Holter monitors, and electrocardiograms (ECG). These tests provide critical cardiac data, such as volumetric heart information (echo), extended period heart rhythms (Holter), and short-term heart rhythms (ECG). The data were anonymised and stored securely in compliance with data protection regulations as part of a retrospective review of anonymised clinical data. Consent was not sought, as the study did not impact patient management.

7.3.2 Data Acquisition and Processing

A total of 49 FD patients and 100 HCM patients were included in the study. Data were manually extracted from the Cardiff and Vale Electronic Clinical Portal, including clinical reports for the three cardiology tests (echo, Holter, or ECG). The reports, stored as .pdfs, contained demographic, quantitative, and qualitative data filled in by cardiothoracic physicians. An automatic document reader tool was used to extract data into a .csv format for analysis.

Only numeric data from these reports was used to train the models. Raw test data, such as ECG waveforms and Echo images, were excluded. Gender and condition were encoded numerically using standard one-hot encoding, and missing data were imputed using zero imputation.

A complete list of features extracted from each test modality (ECG, Echo, Holter) is provided in Table 7.1.

		HCM	Fabry	P-Value
Patients		100	49	
ECG		728	514	
Echo-cardiogram		128	161	
Holter		55	70	
Investigation Periods		**461**	**588**	
Age, mean (SD)		57.2 (13.9)	45.5 (17.4)	< 0.001
Gender, n (%)	Male	367 (62.4)	187 (40.6)	< 0.001
	Female	221 (37.6)	274 (59.4)	
ECG to Echo (Days), mean (SD)		18.0 (21.2)	3.4 (10.3)	0.505
ECG to Holter (Days), mean (SD)		23.9 (11.7)	24.7 (13.3)	0.850

Table 7.1: Summary of the patient data used in the study.

7.3.3 Aligning Clinical Reports

Given this study's limited number of patients, the data were divided into "investigation periods" to create a suitable number of instances for conducting ML. An investigation period was defined as a 45-day window that began from an ECG report in the extracted patient data and included at least that ECG along with any other tests conducted within those 45 days, such as echo or Holter. If multiple tests of the same type occurred within a 45-day period that was not a repeat test from the same day, we separated those into distinct investigation periods. Each investigation period was labelled as having come from an HCM or FD patient and was then

treated as a separate instance in the final dataset. After the data were rearranged in this way, there were 461 investigation periods associated with the HCM patients and 588 with the FD patients (see Table 7.1). This was deemed to reflect the clinical procedures of the University Hospital of Wales, where echo and Holter investigations are commonly conducted in response to an ECG test that the assessing clinician believes requires further investigation. This allowed for the data from the Holter and echo tests to be placed in the context of the ECG assessment that triggered the referral for that investigation.

7.3.4 Modelling Approaches

This work extends the notion of rare event detection to explore an area of machine learning (ML) that deals with the comparative analysis of relatively small sample sizes with high missing values in both case and control groups. Small sample sizes and skewed data increase the likelihood of overfitting and impede generalisation to new patients.

To mitigate this, we tested both individual model optimisation and ensemble approaches. Models included Random Forest, Extra Trees, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) as an optimiser for linear classifiers, GB, eXtreme Gradient Boosting (XGB), and neural networks. SGDClassifier is a linear model that uses SGD as an optimisation method to update model parameters iteratively, making it suitable for large-scale datasets with sparse features. Although SGD is not a model, SGDClassifier is a model that applies SGD for parameter optimisation in classification tasks, such as SVM or logistic regression.

Ensemble techniques leverage these learning algorithms, aggregating predictions to create more robust composites. By combining complementary models, bias and variance are reduced. Specifically, we evaluated stacked ensembles and voting ensembles.

We further optimised ensembles through a cascading process, recursively removing models and re-evaluating them to assess their incremental predictive impact. Models detracting from metrics like the F1 score were eliminated, distilling the ensemble to an optimal combination of complementary contributors that maximised collective accuracy. This empirical ensemble distillation approach extracted core signals from the scarce heterogeneous data.

7.3.5 Statistical Analysis

Following aligning cardiac reports into discrete investigation periods, we conducted a univariate statistical comparison of feature distributions between FD and HCM patients. This pre-

liminary analysis aimed to identify features showing significant group-level differences before model training. Our statistical approach follows the methodology employed by [271].

For each numeric feature, we first applied Levene's test to assess the homogeneity of variances between FD and HCM groups. Levene's test evaluates whether variance differs significantly between groups, informing the choice of subsequent statistical methods. Depending on the outcome of this test, we selected the appropriate follow-up analysis: an independent two-sample *t-test*, suitable when variances between groups are approximately equal and the assumption of normality holds, or the non-parametric *Mann–Whitney U test* (also known as the Wilcoxon rank-sum test), appropriate when equal variances or normality assumptions are not satisfied. To quantify the magnitude of these group differences, we calculated effect sizes using Cohen's d for results derived from t-tests and rank-biserial correlation (r) for those from Mann–Whitney U tests. Features with statistically significant differences ($p < 0.05$) were retained for further visualisation and interpretation.

7.3.6 Cross-Validation and Hyperparameter Optimisation

We also employed repeated stratified K-fold cross-validation. Cross-validation reduces reliance on individual splits by creating multiple train-test splits, preserving case/control distributions. It provides tighter estimates of real-world F1 performance compared to standard train/test evaluation. We performed a 10-fold CV, repeated three times with different splits. F1 scores were averaged across all folds and repetitions to minimise variance and overfitting risks posed by limited clinical data.

Tuning model configurations is critical for preventing overfitting with limited samples [272]. We leveraged Bayesian optimisation methods, including Optuna, to efficiently find well-suited hyperparameter combinations that maximised the repeated stratified K-fold cross-validated F1 scores. Despite scarce, heterogeneous clinical time series data, the optimisations produced a generalisable model.

7.3.7 Explainability

Clinically validating the model's decisions and thresholds was central to our approach. The techniques chosen below provide a coherent exploration of the model's decisions to enlighten clinical practice.

7.3.7.1 Feature Importance

ML models like XGB, random forests, and other tree-based methods have built-in feature importance calculations that score each input's contribution to predictions. The calculation involves several steps. Firstly, the "Split Gain" measures the impact of a feature when splitting decision tree nodes, assigning higher importance to features leading to a more substantial reduction in the loss function. Secondly, "Cover" sums the relative quantity of data points within a feature, emphasising features covering a larger portion of the dataset. "Frequency of Feature Use" tracks how often each feature is utilised in decision-making across all ensemble trees, prioritising features frequently used for splitting nodes. The final importance score sums each feature's split gains, cover values, and frequency of feature use, expressing these as a proportion of the overall feature importance.

7.3.7.2 Permutation Feature Importance

In addition to the built-in feature importance, we also assessed permutation feature importance. This technique evaluates a feature's impact by randomly shuffling its values and measuring the resulting decrease in the model's performance. Features for which shuffling values significantly decreases metrics like AUROC are deemed highly important. By focusing on key signals in the small, heterogeneous dataset, this method helps corroborate other explanations.

7.3.7.3 SHAP Feature Importance

SHAP values offer a detailed exploration of each feature's contribution, making them particularly useful for complex models like XGB [126]. Rooted in cooperative game theory, SHAP fairly distributes "credit" for a prediction among features by evaluating the impact of all possible feature combinations. This method generates comprehensive insights into global and instance-level influences, addressing the limitations of built-in feature-importance techniques. By revealing interactions and directionality that often go undetected in traditional approaches, SHAP provides validated explanations for differentiating patterns within heterogeneous clinical data.

7.3.7.4 Partial Dependence Plots

While feature importance scores provide a ranking of predictive influences, Partial Dependency Plotss (PDPs) offer a more detailed view of how changes in a variable impact predic-

tions. PDPs illustrate the marginal effect of a feature on predicted outcomes by averaging model predictions across samples as the feature is varied. By creating visualisations of these effects across the feature's value range, even in non-linear models like neural networks, PDPs help identify thresholds and reasonable value ranges for clinical variables. These insights can inform personalised assessment and treatment decisions. We consulted with domain experts for accurate interpretations in the analysis. Integrating multiple explanation techniques was key to obtaining reliable insights, even with limited heterogeneous data.

7.4 Results and Evaluation

7.4.1 Data Distribution

There were twice as many HCM patients (N=100) as FD patients (N=49). Compared to FD patients, HCM patients were older (mean age of HCM patients was 57 and 45 for FD patients) and had more males (HCM males: 62%, females 38%; FD males: 40%, females: 60%). These differences may have resulted from genetic and biological factors driving divergent disease onset and severity between sexes [273].

Regarding aligned cardiac tests within investigation periods, a significantly lower 3-day average delay between ECG and Echo for Fabry patients indicates more rapid multimodal profiling compared to the 18-day average lag in HCM patients. However, comparable uniform Holter intervals with 24-day lags between ECG and Holter for both diseases suggest consistency on broader test alignment timescales between cohorts.

7.4.2 Statistical Analysis

To complement the model-driven analyses, we first assessed univariate differences between FD and HCM patients across all numeric features. Of the total features tested, 24 were statistically significant ($p < 0.05$). Figure 7.1 displays the corresponding effect sizes for these significant features.

On average, features with positive effect sizes were higher in FD patients, while negative values indicate higher means in HCM patients. T-axis, QTc, and age emerged as the most distinguishing features favouring Fabry classification, while Left Ventricular Pulse Wave Duration (LVPWd) and EDV (MOD-sp2) were more associated with HCM cases. These findings offer an initial statistical grounding and are later compared with the features prioritised by machine learning interpretability techniques.

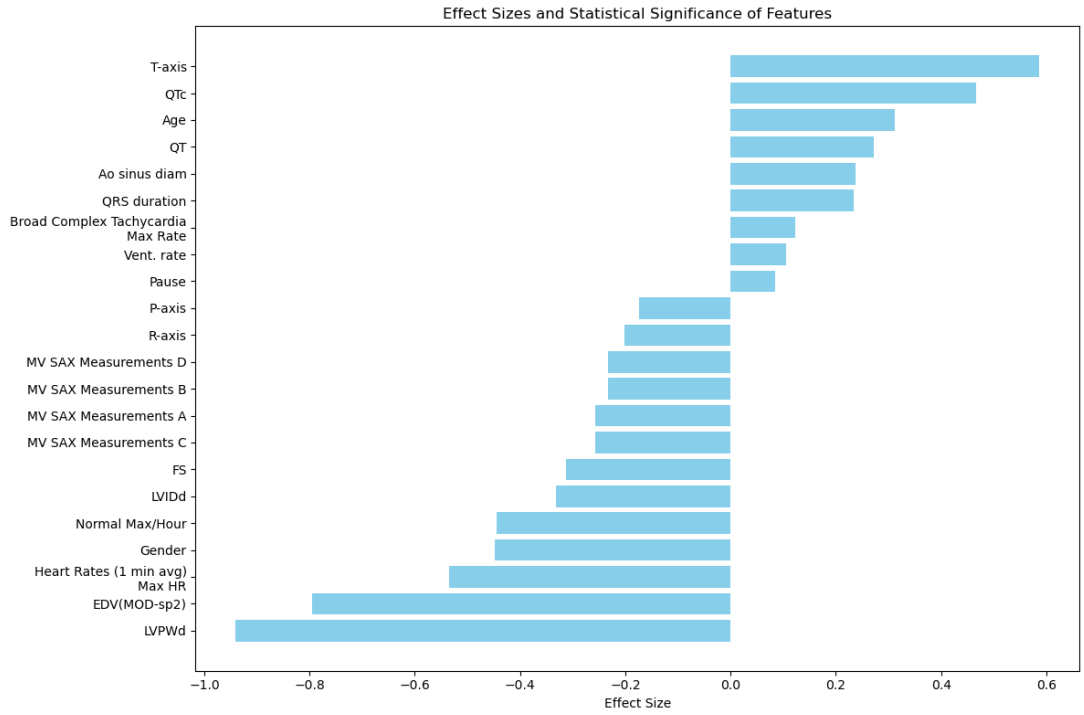


Figure 7.1: Effect sizes of statistically significant features comparing FD and HCM groups. Positive values indicate higher values in FD patients.

7.4.3 Model Evaluation

Model performance was evaluated on held-out test data using key classification metrics as shown in Table 7.2. Recursively eliminating underperforming methods resulted in a final voting ensemble that comprised a group of tree-based and distance-based models, XGB, K-Nearest Neighbors (KNN) and Endothelin (ET). The final ensemble achieved an F1 score of 0.90. The stacking ensemble consisted of non-parametric, gradient-optimised learning models KNN, SGDClassifier, and Multi-Layer Perceptron (MLP), which achieved a slightly better F1 score of 0.91. Both ensemble methods performed well but did not surpass the individual XGB classifier, with an F1 score and AUC of 0.92.

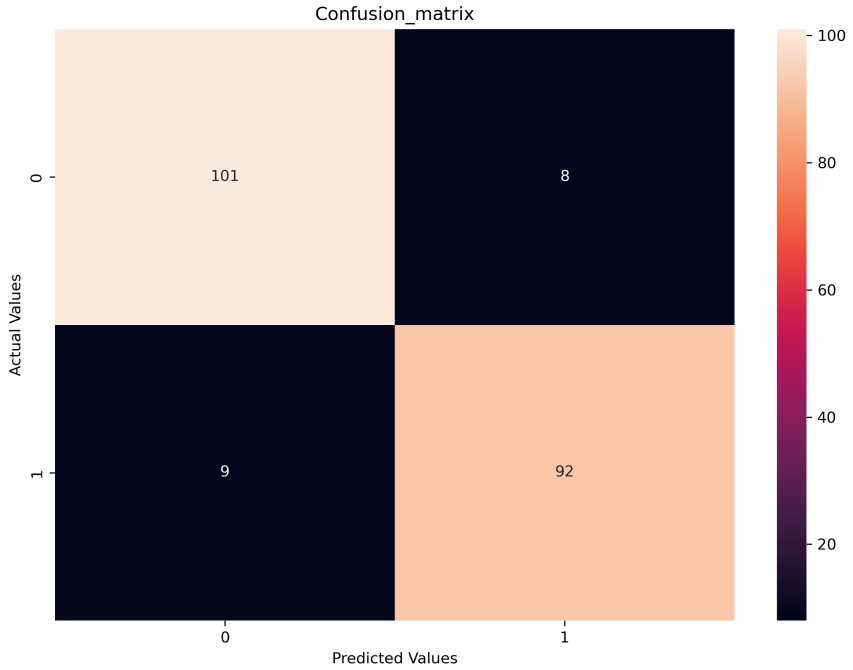


Figure 7.2: Confusion matrix of optimised XGB classifier for FD (1) and HCM (0).

While tree-based models tend to work better for this dataset, substantial gains from gradient boosting and tuning XGB highlight the effectiveness of this approach in discovering complex, rare disease patterns. Additionally, the tendency towards predicting Fabry disease and the lower False Positive Rate (FPR) of 0.0734 vs a False Negative Rate (FNR) of 0.0891, highlights the models has a high clinical utility at prompting patients suspicious of FD for genetic testing (see 7.2).

7.4.4 Feature Importance

Fig. 7.3 shows the most predictive marker overall as Left Ventricular Internal Dimension at Systole (LVIDs), which indicates sensitivity to gross structural reshaping. Complementing this are the Interventricular Septum at End-Diastole (IVSd) wall thickness and Left Ventricular Internal Dimension at Diastole (LVIDd) dimension, which together quantify the chamber anatomy rearrangements through hypertrophy or glycogen infiltration.

Conduction metrics formed the next highest cluster, including repolarisation durations (QTc) and depolarisation wavefront aberrations (T-axis), along with general speed (QRS du-

Model Name	Accuracy	Precision	Recall	F1 Score
Decision Tree (DT)	0.80	0.80	0.80	0.80
Random Forest (RF)	0.88	0.88	0.88	0.88
Extra Trees (ET)	0.81	0.82	0.81	0.81
K-Nearest Neighbors (KNN)	0.81	0.81	0.81	0.81
SGDClassifier	0.64	0.69	0.64	0.60
Naive Bayes (NB)	0.55	0.62	0.55	0.46
Logistic Regression (LR)	0.73	0.73	0.73	0.73
AdaBoost	0.80	0.81	0.80	0.80
Gradient Boosting (GB)	0.89	0.89	0.89	0.89
Extreme Gradient Boosting (XGB)	**0.92**	**0.92**	**0.92**	**0.92**
Multilayer Perceptron (MLP)	0.74	0.76	0.74	0.74
Stacking Ensemble	0.91	0.91	0.91	0.91
Voting Ensemble	0.90	0.90	0.90	0.90

Table 7.2: Test set performance of top models.

ration) and directionality (R-axis, P-axis) measurements. Together, these biomarkers track functional excitation-contraction coordination.

Lastly, other influential markers include patient gender, baseline heart rate tendencies (bradycardia), arrhythmia frequency metrics and relative age-of-onset with demographic factors likely related to risk and progression elements influencing disease manifestation.

7.4.5 Permutation Feature Importance

Beyond assessing feature strength, permutation analysis reveals the model's reliance on specific inputs by quantifying its performance when primary features are altered. Stress-testing each variable by shuffling its values highlights the adaptable markers that anchor decisions amidst uncertainty.

The T-axis stands out in the permutation analysis (refer to Figure 7.4), showing a pronounced reliance on the morphology of repolarisation waves for distinguishing between conditions when other cardiac indicators are varied. This aligns with previous findings that cardiovascular health is a predictor of Fabry disease [274]. Both age and gender also emerge as key variables, consistent with existing literature.

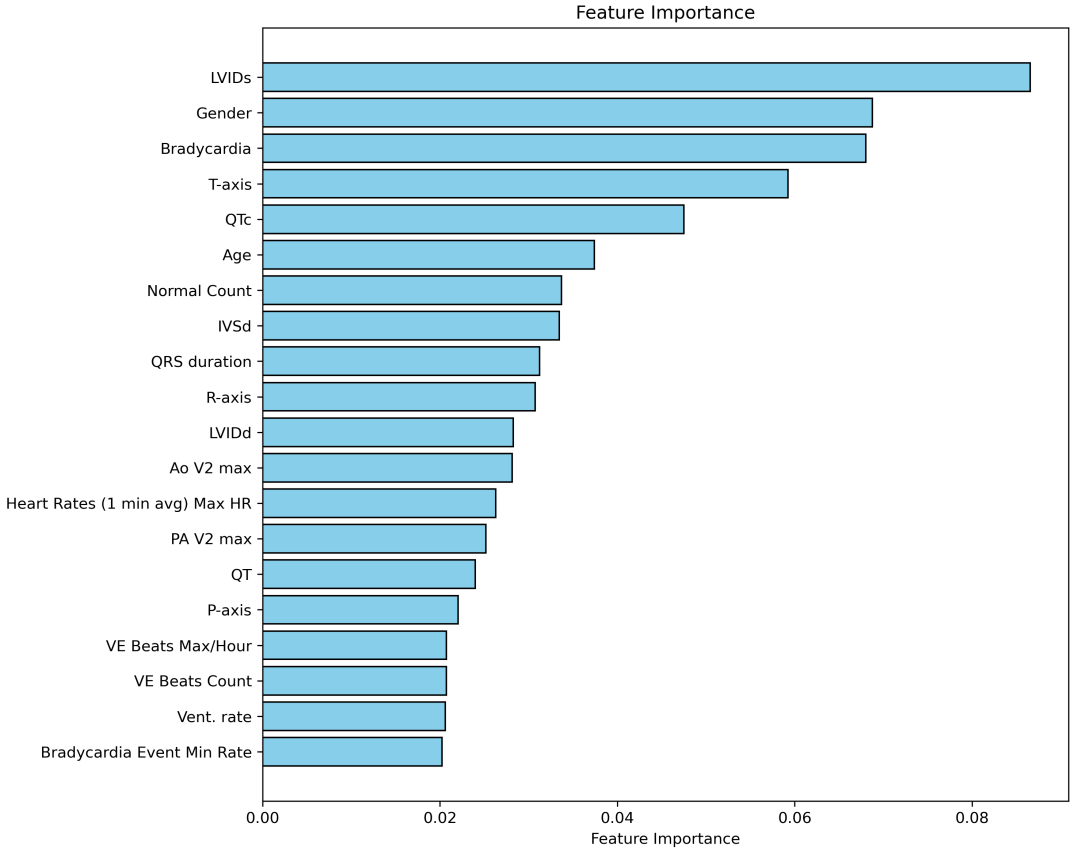


Figure 7.3: Top 20 most important features of the XGBoost model ranked in order of importance

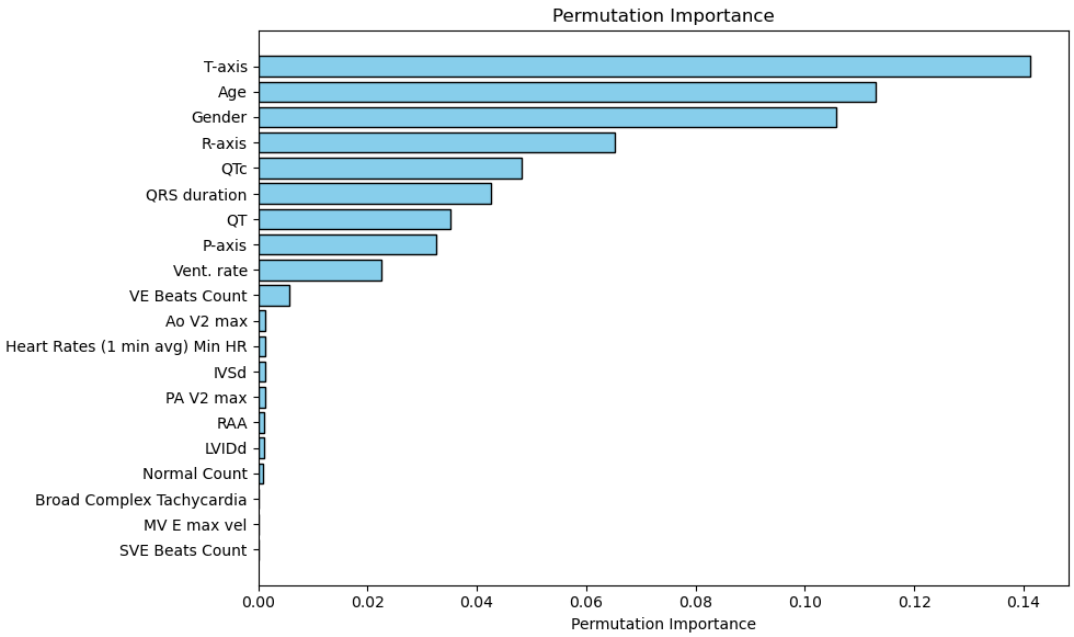


Figure 7.4: Bar graph of top 20 influential features in model prediction.

The persistence of factors like conduction speed (QRS duration), the heart's electrical conductivity (R-axis), and the timing of electrical recovery (QT/QTc) suggests that complex wave patterns are essential for diagnosis.

Other significant indicators included the frequency of irregular heartbeats (PVC counts), the speed of blood flow in the heart (Ao/PA peaks), and general heart rate indicators.

7.4.6 Shapley Feature Importance

7.5 shows that lower values for age and T-axis were more associated with FD classification, while higher values for gender (where 1 = male and 2 = female) were associated with FD (so the model was more likely to assign a patient to FD if they were female).

QTc length only pushes predictions strongly toward abnormal extremes, acting as a tipping point for some. Meanwhile, the ventricular rate maintains a fairly balanced directionality, indicating less differential impact.

The multi-faceted perspective of SHAP analysis emphasises subtle and overt feature influences lost in summary statistical approaches.

7.4.7 Partial Dependence Plots

Partial dependence plots characterise how subtle changes in factors alter predicted condition probabilities.

The relationship with age (Figure 7.7) showed that younger ages (especially under 38) were associated with increased FD likelihood. At the same time, predictions skewed toward HCM with growing age, especially beyond 65 years. Gender showed consistent likelihood differences across ages, suggesting prolonged progression divergence.

Cardiac electrical markers also presented deterministic thresholds. Figure 7.6 indicates T-axis patterns temporarily boosted FD likelihood up to 45 ms before predictions swung towards HCM, reflecting potential changes in severity levels over time. QTc interval predictions were consistent until 320, where the likelihood of FD increased slightly, but at 400 ms to 480 ms, a step decline indicated a higher likelihood of HCM (Figure 7.8).

Additional factors like R-axis and P-axis exhibited oscillating relevance dips and spikes. Functional elements demonstrated further delineations, with QRS duration increasing HCM likelihood between 80-100 ms thresholds and then increasing rapidly after 100 ms, signalling a higher risk of FD. While faster maximum heart rates increased the likelihood of HCM, metrics like ectopy counts presented a high likelihood of FD between 0 and 3 beats.

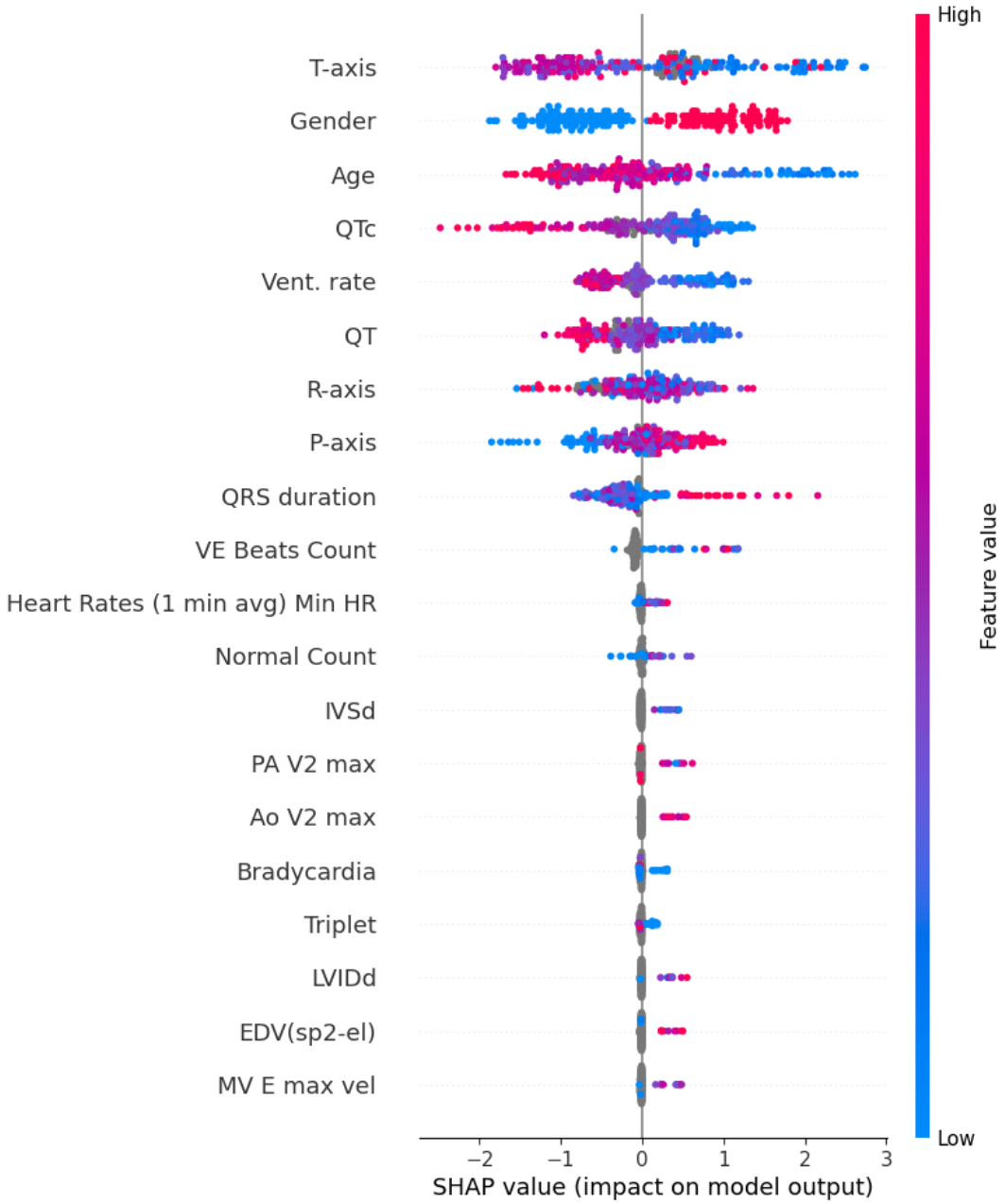
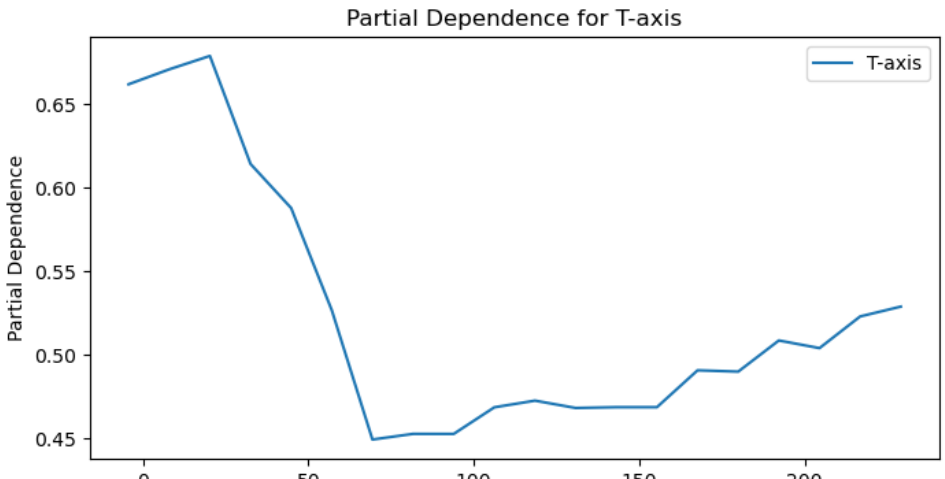


Figure 7.5: SHAP Feature Importance for top 20 features



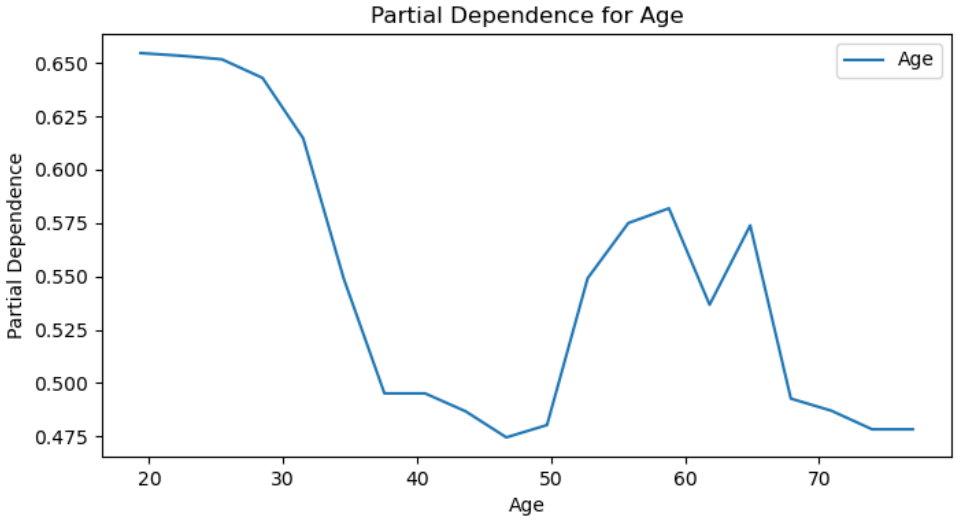


Figure 7.7: Partial dependence plot illustrating age impact on FD and HCM prediction.

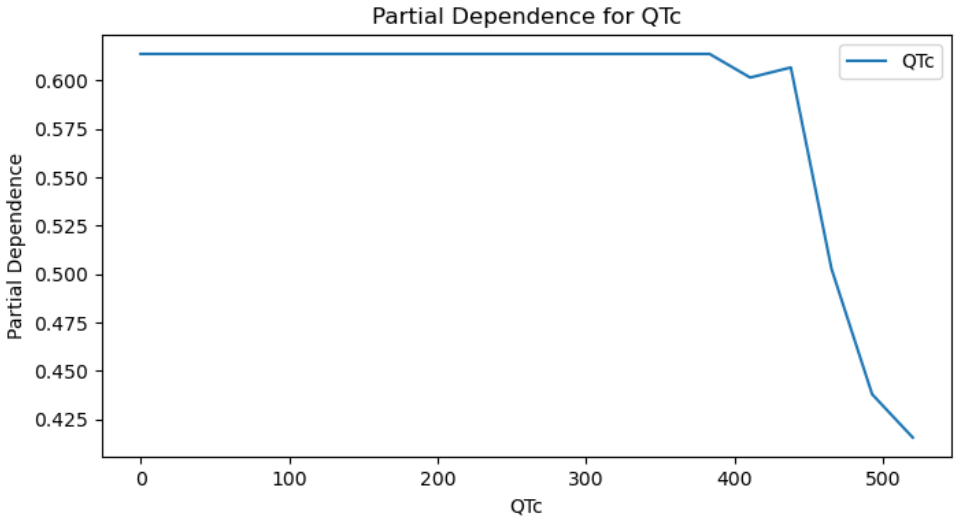


Figure 7.8: Partial dependence plot showing QTc interval's effect on case differentiation.

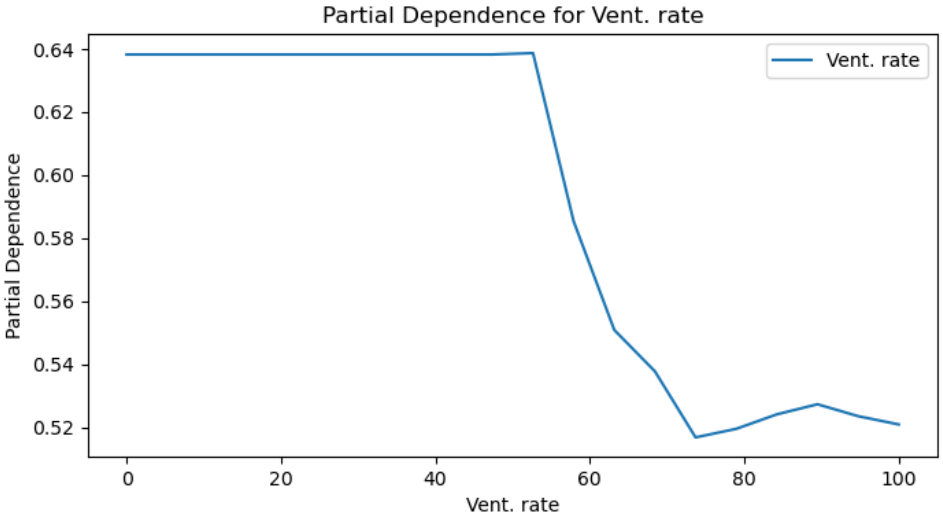


Figure 7.9: Partial dependence plot for Ventricular Rate's role in predictions.

7.5 Discussion

FD is a rare condition that has long been understudied due to limited research and the scarcity of data. The symptoms of FD can often be mistaken for those of more common conditions, leading to significant diagnostic delays—sometimes lasting up to 15 years—while patients are shuffled between specialists who rule out more prevalent diseases [256, 257]. During this time, irreversible complications such as neuropathy, nephropathy, and cardiomyopathy can accumulate without being detected [253].

Although some studies have identified markers that differentiate FD from other conditions, the choice of control groups in these studies often fails to reflect the real-world challenge of misdiagnosis, particularly concerning diseases like Hypertrophic Cardiomyopathy (HCM), which shares several clinical features with FD. The difficulty of distinguishing between these two conditions presents a vital opportunity for improving diagnostic efficiency. Furthermore, previous studies have been constrained by small cohort sizes and a lack of comprehensive interpretability of the models, contributing to clinicians' reluctance to apply these models in clinical practice.

This study addresses these challenges by utilising a larger cohort and incorporating multisystem interpretability techniques such as SHAP to offer a transparent explanation of the model's decision-making process. We aim to solve the challenge of distinguishing between

FD and HCM by employing multimodal cardiac data (ECG, Echo, and Holter) and machine learning. By aligning these diverse data sources into longitudinal investigation periods, we reveal phenotypic differences between FD and HCM that were previously difficult to detect [262, 275, 276]. Using interpretable AI models improves upon earlier efforts, such as electrocardiogram-based models, by offering clearer insights into why the model predicts one condition over another, ensuring clinicians can trust and apply the model's findings confidently.

7.5.1 Comparing Approaches

The algorithm we developed synthesises structural, electrical, and demographic data from multimodal patient records, achieving an F1 score and AUC of 0.92 in differentiating between FD and HCM.

The model reveals significant influences from subtle tissue-level shifts, such as T-axis waveform deviations, which indicate early depolarisation disruptions—an early FD marker often overlooked in conventional electrocardiographic analyses. Additionally, age and gender cutoffs provide insights into demographic factors correlating with disease risk, reinforcing clinical understanding of FD behaviour. Notably, this model's findings go beyond current clinical guidelines, which do not set explicit thresholds for suspecting FD or HCM in patients.

Our model improves upon Moura *et al.* 's work, which used electrocardiogram data to achieve an AUC of 0.87 in differentiating FD from HCM [262]. While their model highlights disrupted depolarisation and excessive recovery risks, it was limited by low specificity (56%) and a small, age-stratified cohort of just 64 patients.

The use of broader medical records in our model enables us to explore arrhythmia and stroke risks, as seen in studies [275, 276], which achieved an AUC ranging from 0.77 to 0.81. However, these studies heavily rely on observational proxies such as palpitations or previously confirmed events, limiting their ability to personalise risk stratification.

7.5.2 Clinical Insight

Our deep dive into the model's decision-making process has uncovered critical patterns that enhance our understanding of FD and provide early indicators that can be applied in clinical practice. These patterns, derived from the model's interpretable techniques, allow for a more nuanced approach to identifying risk factors long before noticeable disruptions in heart rhythm occur.

One of the most prominent markers we identified is the T-axis, a key component in the electrocardiogram (ECG) waveform. When analysed through techniques like SHAP and partial dependence plots, this feature reveals that the likelihood of FD increases within specific 0-25 millisecond aberrancy ranges. These deviations likely signal early electrical propagation delays in the Purkinje fibres, which are essential for rapid conduction in the heart. These fibres are damaged in FD patients due to glycosphingolipid accumulations, a disease hallmark. However, these Purkinje fibres remain intact in patients with Hypertrophic Cardiomyopathy (HCM), underscoring one of the key differences between the two conditions.

Alongside structural changes in the heart, such as ventricular hypertrophy, we observe timing disruptions within the ECG readings that point to cellular-level conduction issues emerging early in FD progression. Techniques like SHAP values and permutation feature importance help illuminate these disruptions, providing insights into the QTc interval—a measure of the heart's electrical activity. Our analysis reveals that FD patients diverge at the 400 ms QTc threshold, indicating higher instability risks. In contrast, patients with HCM show minimal disruption at this threshold, reinforcing that these two conditions share overlapping symptoms and can be distinguished with nuanced electrical markers.

The use of many interpretability techniques, such as SHAP and permutation feature importance, has allowed us to go beyond just identifying important features. It has enabled us to understand how these features influence the model's prediction for example, SHAP values provide detailed insights into the individual contributions of each feature, such as gender, age, and cardiac markers, to the likelihood of a diagnosis. This deep understanding of feature importance allows clinicians to see which factors influence the model and validate them against their clinical experience.

Moreover, PDPs visually represent how changes in specific variables—such as T-axis measurements or QTc interval values—directly affect the model's predictions. This helps clinicians understand the thresholds at which key markers shift, making tracking early signs of FD and HCM easier. For instance, the T-axis pattern, which temporarily boosts the likelihood of FD within a narrow window, serves as an early diagnostic cue for FD, which can be distinguished from HCM. The QTc interval divergence at the 400 ms threshold similarly marks a critical distinction between the two conditions, helping clinicians identify FD earlier in the diagnostic process.

This interpretability framework provides actionable insights that can be directly integrated into clinical practice. Clinicians can now track key markers—such as T-axis variations, QTc

divergence, and age—to identify early signs of FD, improving personalisation of care. These markers offer early diagnostic windows that allow clinicians to intervene before irreversible organ damage occurs. By personalising care based on these findings, clinicians can make timely, informed decisions that may significantly improve patient outcomes.

Moreover, the SHAP values allow clinicians to understand the model's predictions in real-time, providing transparency and trust. When clinicians have a clearer view of why the model makes specific predictions, they can confidently apply the findings. This explainable AI (XAI) approach fosters trust in the model and encourages its use in routine clinical practice, enabling early detection and differentiation between FD and HCM.

In summary, leveraging multisystem interpretability techniques in this study has enhanced the model's accuracy and provided clinically actionable insights. By focusing on nuanced features such as T-axis variations and QTc interval divergences, we can now differentiate between FD and HCM more effectively, empowering clinicians to diagnose earlier, personalise care, and ultimately improve patient outcomes.

7.5.3 Limitations and Future Directions

Despite the multisystem interpretability framework enhancing model reliability, several limitations remain. These include the retrospective study design, potential selection bias in control groups, and challenges related to the small sample size. Future work could address these issues by evaluating additional cardiac data modalities, such as genetic testing for variants in MYH7, MYBPC3, and GLA genes. This may offer more molecular insights into the observed electrical and anatomical patterns.

One limitation of identifying undiagnosed FD from clinical populations is that the condition often reaches a symptomatic stage before detection, as highlighted by Jeffries *et al.* [260]. The challenge remains to diagnose FD before irreversible damage occurs. However, as noted earlier,

7.6 Summary

This work advances ML applications in healthcare by designing a robust XGB classifier to differentiate between FD and HCM—a common misdiagnosis for FD. Our model leverages routine clinical data (ECG, Holter, echocardiograms) already collected in clinical settings, making integrating it into real-world healthcare workflows easier. This marks a significant improve-

ment over previous work, where ML models typically struggled to work with routine data and required complex or rare datasets.

One of the key contributions of this study is the creation of a novel dataset explicitly tailored for clinicians and researchers to use in diagnosing rare diseases like FD. This dataset, derived from real-world clinical data, not only helps researchers build better models but also serves as a valuable resource for education and further research in rare disease diagnosis.

In addition, we developed a custom data extraction tool that automates the process of extracting clinical data from non-standard report formats. This tool, which has already been tested and proven useful across multiple health boards, simplifies data integration from legacy systems, making it easier to work with fragmented data and improving data accessibility for model training.

Our XGB classifier is designed to work with clinically collected data and can differentiate FD from HCM with an impressive F1 score of 0.92. The model has been built with explainability in mind, using techniques like SHAP values and PDPs to provide transparent, interpretable predictions that clinicians can trust and easily integrate into their decision-making processes.

This approach represents a significant step beyond previous work on sepsis, where we focused primarily on understanding missingness patterns in ICU data and modelling them using XGB. While the sepsis work also used SHAP for interpretability, it was primarily centred around feature selection and data imputation, with a more limited focus on clinical applicability. In contrast, this study offers a deeper clinical insight by identifying key features like T-axis variations and QTc interval deviations and making these insights actionable in a real-world clinical context.

In summary, this work demonstrates the potential of ML to support clinicians in diagnosing rare diseases like FD by providing a clinically applicable model that integrates routinely collected data, improves diagnostic accuracy, and enhances clinical decision-making. Adding the new dataset and data extraction tool further solidifies the clinical utility of this work, making it easier to implement in real healthcare environments. Future work will focus on comparing the effectiveness of this classifier against expert clinicians to assess its impact on diagnostic speed and accuracy in clinical cardiology settings.

7.7 Connecting the dots and remembering the why

As we advance in applying ML to diagnosing rare diseases, we must stay grounded in the "why" behind these efforts. At the heart of this research is a commitment to improving patient

outcomes by enabling earlier and more accurate diagnoses. The goal is not just to identify rare diseases more efficiently, but to personalise treatment plans based on each patient's unique clinical profile, considering the disease and its broader context within their health journey.

In the case of FD, which has long been an understudied condition, our work serves as a beacon for the potential of ML in clinical practice. By providing clinicians with tools that distinguish rare diseases from common conditions and explain why these distinctions matter, we empower them to make more informed decisions. This can lead to faster diagnosis, better-targeted treatments, and, most importantly, improved patient care.

The long-term vision is to ensure that every patient, regardless of how rare their condition, receives the timely diagnosis and tailored care they deserve. Through tools like the XGB classifier and its interpretable features, we hope to bridge the gap between rare and common conditions, fostering a healthcare ecosystem that recognises the complexity of all diseases and uses ML to support clinical decision-making every step of the way.

Chapter 8

Augmenting Machine Learning predictions with LLMs

Contents

8.1 Introduction

ML models have made significant strides in identifying complex patterns within medical datasets, such as imaging, genomics, and patient records. In recent years, MLs ability to assist clinicians in diagnosing conditions with unprecedented speed and accuracy has been widely acknowledged. However, rare diseases such as FD present unique challenges. FD is characterised by its heterogeneous clinical presentation, making it difficult to distinguish from other more common conditions, such as HCM, which shares similar cardiac symptoms. This often leads to misdiagnosis or delays in diagnosis, contributing to severe complications like irreversible organ damage and reduced quality of life for patients. Additionally, its rarity of FD means

that clinicians have limited exposure to it, further compounding the difficulty in recognising its early signs.

Incorporating ML into diagnosing rare diseases like FD holds great promise, but challenges remain, especially in real-world clinical integration. The lack of training data, model transparency, and trust from clinicians is a major obstacle to the widespread adoption of AI in healthcare.

This chapter addresses these challenges by proposing an innovative ML-LLM diagnostic tool that integrates the FD-HCM ML model with an LLM-based chatbot to provide clear, clinically relevant explanations. This tool provides accurate diagnostic predictions and aims to enhance transparency and interpretability, making it easier for clinicians to understand and trust AI-driven decisions.

FD is a rare, genetically inherited metabolic disorder that affects approximately 1 in 15,000 people worldwide. It is caused by a deficiency in the alpha-galactosidase A enzyme due to mutations in the GLA gene, leading to progressive damage to the nervous system, kidneys, and heart. The clinical presentation of FD is highly heterogeneous, with symptoms ranging from chronic pain, fatigue, and clouded vision to severe cardiovascular events such as arrhythmias, strokes, and kidney failure. These symptoms are often intermittent and mild, making it difficult for clinicians to recognise the disease early. Furthermore, many of these symptoms overlap with those of more common diseases, particularly HCM, a condition that shares similar cardiac symptoms, such as left ventricular hypertrophy and arrhythmias.

Diagnosing FD is challenged by its rarity, heterogeneous presentation, and overlapping symptoms with common diseases. These factors contribute to misdiagnosis and delayed treatment, with patients often facing diagnostic delays ranging from 3 to 15 years. This delay in diagnosis has significant consequences, leading to irreversible organ damage (particularly in the kidneys and heart), reduced quality of life, and increased healthcare costs.

Key complexities of FD diagnosis include:

- **Rarity of the Condition**: As a rare disease, FD is not commonly encountered in clinical practice. Clinicians, therefore, have limited exposure to the condition, making it harder to recognise and diagnose early.

- **Overlapping Symptoms with Common Diseases**: Many FD symptoms overlap with more common diseases, such as HCM, coronary artery disease, and chronic fatigue syn-

drome. This often leads clinicians to prioritise these more common conditions during the diagnostic process, delaying the identification of FD.

- **Heterogeneous Clinical Presentation**: The symptoms of FD can vary greatly between individuals. Some patients may exhibit severe symptoms early in life, while others may experience more subtle or intermittent symptoms that are easily misattributed to other conditions. This variability complicates the application of standardised diagnostic protocols.

- **Sporadic Clinical Encounters**: Due to the rarity of the disease, FD is often not considered in the differential diagnosis until other more common causes have been ruled out, leading to significant delays in diagnosis and treatment.

- **Data Scarcity**: FD is a poorly studied disease due to its rarity, and consequently, there is limited data available to support the development of diagnostic tools. The lack of sufficient training data complicates the development of practical diagnostic algorithms.

These complexities highlight the necessity of the approach presented in this chapter. Despite the promise of ML in diagnosing rare diseases, a key barrier to its widespread clinical adoption is the lack of interpretability. Many ML models function as "black boxes," offering little insight into how predictions are made. This lack of transparency is particularly problematic in the medical field, where clinicians must trust the model's reasoning before applying it to patient care.

To address this, we propose an innovative approach that integrates the FD-HCM ML model with an LLM chatbot. This integration is designed to provide accurate predictions and offer clear, actionable explanations of the model's reasoning. The use of LLMs to augment ML predictions addresses the transparency issue by enabling the model to generate natural language explanations, making its decision-making process more understandable to clinicians.

This chapter describes the development of the ML-LLM diagnostic tool, which aims to assist clinicians in the early diagnosis of FD by providing accurate predictions and a transparent rationale behind those predictions. We also explore the clinical co-design approach to ensure the tool's relevance and usability in clinical settings. By improving the interpretability and usability of the model, this tool has the potential to enhance the accuracy and speed of FD diagnosis, ultimately improving patient outcomes through earlier interventions.

The following sections discuss the design considerations for integrating the ML model with the LLM chatbot, including user interface design, clinical workflow integration, and evaluation

of the prototype tool through a user study. We will also address the limitations, challenges, and future directions for integrating ML and LLM technologies into clinical practice for rare disease diagnosis.

The tool developed in this chapter can be accessed here: **AI4Fabry Diagnostic Tool**.

8.2 Designing Systems for Clinical Interpretability

8.2.1 Prompt Engineering for Clinically Relevant Explanations

Prompt engineering is designing and optimising the queries to guide LLM in generating clinically relevant outputs [277]. Effective prompts balance the open-ended nature of natural language and the specificity required in medical contexts. They should be clear, concise, and structured to elicit responses that apply to the patient's situation, incorporating relevant medical terminology and context [278].

Developing high-quality prompts is an iterative process that relies on clinician feedback. Initial prompts are tested, and the LLM's outputs are evaluated by healthcare professionals who highlight areas for improvement in clarity, relevance, and accuracy. This process is repeated until a set of optimised prompts is obtained. This collaborative feedback loop ensures that the LLMs responses are aligned with clinical needs, merging the technical expertise of engineers with the practical, patient-centred experience of clinicians.

In the context of our decision support tool for FD, prompt engineering will help ensure that the chatbot explains selected features and its decision-making process in a clinically relevant manner and can be easily understood by healthcare professionals.

8.2.2 Designing Intuitive User Interfaces

The user interface (UI) serves as a bridge between complex data algorithms and healthcare professionals. Its design should prioritise ease of use and efficiency. A well-designed interface reduces cognitive load, allowing clinicians to access and interpret model outputs seamlessly while focusing on patient care. Key considerations include clear data visualisation, logical information organisation, and streamlined workflows that mirror clinical processes. By enhancing the clinician's ability to utilise the LLMs insights, an intuitive interface facilitates the effective integration of AI-assisted tools into clinical practice.

Our diagnostic tool's UI has been designed with these principles in mind, focusing on providing a seamless and intuitive experience for clinicians. The interface incorporates clear

visualisations of the ML model's predictions and the SHAP values, allowing clinicians to grasp the key factors influencing the diagnostic suggestions promptly. The workflow has been optimised to align with clinical processes, ensuring that the tool can be easily integrated into existing diagnostic protocols.

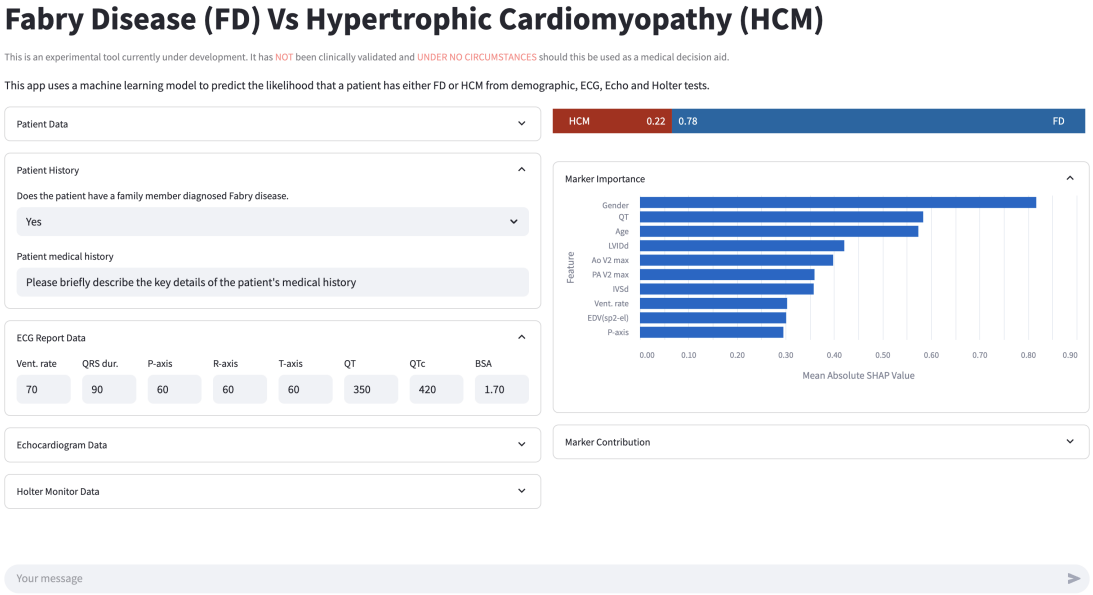


Figure 8.1: Full diagnostic tool interface showcasing the patient data input sections, ECG report data, risk prediction results, and marker importance.

8.2.3 Developing a User-Friendly Streamlit Interface for Clinicians

Building systems on publicly available tools promotes the democratisation of healthcare technology and broadens its impact [279]. Open-source development encourages collaboration, validation, and improvement of AI-assisted diagnostic tools across healthcare settings. It facilitates the standardisation of care quality, regardless of an institution's resources, and fosters innovation through the widespread adoption of best practices [280].

Our diagnostic tool has been developed using the open-source framework Streamlit, which is freely available for anyone to use and modify. Streamlit is a lightweight and user-friendly framework for creating web applications for data science. It has a clean and uncluttered layout, apt for clinical decision-making, as shown in Figure 8.1. The simple and efficient interface would allow clinicians to input patient data and receive risk assessments with minimal training or technical knowledge.

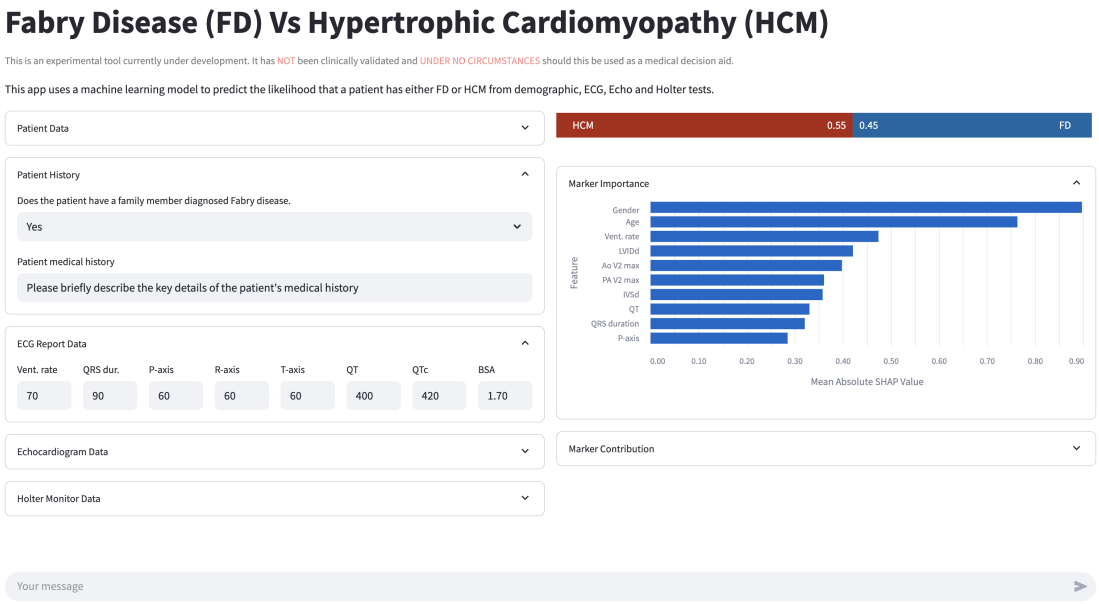


Figure 8.2: Full diagnostic tool interface showcasing the patient data input sections, ECG report data, and the risk prediction results, along with marker importance with modified risk to showcase change in interface from Figure 8.1

Further, Streamlit's accessibility helps ensure the broader healthcare community can easily share and improve our code and methodologies. The complete source code for our web application is hosted on GitHub, a popular platform for open-source software development and collaboration [281]. By making our code publicly available on GitHub, we encourage other researchers and developers to build upon our work and create similar tools for different diseases.

Additionally, our diagnostic tool utilises state-of-the-art language models streamed from free instances on the Hugging Face Hub [282]. Hugging Face is a leading platform for open-source ML models, providing a centralised repository where researchers can share their pre-trained models with the community. Integrating these freely available models into our tool ensures that users can access cutting-edge AI capabilities without incurring additional costs.

8.2.4 Ensuring Compliance with Data Privacy Regulations

We have implemented several mechanisms within the diagnostic tool to ensure compliance with data privacy regulations. As this is a smart risk calculator, no queries or values input by users are stored. The interface is specifically designed for manual input only, and the system does not allow uploading identifiable patient files. This approach ensures that no sensitive patient data is

retained, processed, or transmitted, mitigating potential privacy risks. Additionally, the system fully complies with standard data privacy guidelines, including the General Data Protection Regulation (GDPR).

8.3 Enhancing Clinicians' Understanding of Model Predictions

8.3.1 Providing Clear Visualisations of the Model's Predictions and Explanations

Clear visualisations are essential for clinicians to quickly and effectively understand the predictions and explanations of ML models. Visualisations allow clinicians to grasp the most important aspects of the data and models by simplifying complex concepts and highlighting key insights. Our diagnostic tool employs SHAP values and LIME to show how input features affect model predictions, making the decision-making process transparent.

By integrating these visualisations with tools like MLflow, we can ensure that they correspond to specific model runs and states, providing consistency and accessibility for clinicians. This integration also offers benefits such as persistent storage and provenance, which are crucial for maintaining the integrity of the model analysis.

8.3.2 Enhancing Transparency with SHAP Values

SHAP values serve to allow for transparency by quantifying how each feature in a patient's data contributes to the diagnostic suggestion or risk assessment from the LLMs output [126]. By visualising SHAP values through bar plots or heatmaps, clinicians can quickly identify the most influential factors driving the model's predictions. This information can guide further investigation and help clinicians make more informed decisions about patient care [283].

In addition to feature importance, visualising the data preprocessing steps through flowcharts or diagrams can help clinicians understand the transformations the data undergoes before being fed into the model. This transparency builds trust in the system and allows clinicians to assess the validity of the model's inputs.

8.3.3 Enabling Clinicians to Input Patient Data Seamlessly and Explore Different Scenarios

We have designed an interface that allows clinicians to input patient data seamlessly and explore different patient scenarios by adjusting input variables, providing valuable insights into how changes in patient data might affect the model's predictions. This interactive exploration helps clinicians understand potential outcomes and prepare for adverse clinical situations.

Our interface incorporates visual tools like partial dependence plots, illustrating how specific features influence predictions. These interactions allow clinicians to understand the model's logic better and perform scenario analyses. For instance, a clinician could adjust the values of specific biomarkers or imaging features to see how the model's risk assessment for Fabry disease changes, providing a more comprehensive view of the patient's condition.

8.3.4 Utilising the LLM Chatbot to Articulate Decision-Making Processes in a Conversational Manner

One of our diagnostic tool's key features is integrating a chatbot powered by an LLM. This chatbot serves as an interactive guide, articulating the decision-making processes of the ML model conversationally. By engaging with the chatbot, clinicians can ask questions and receive explanations in natural language, making the information more accessible and understandable.

For instance, a clinician could ask the chatbot why a patient is classified as high-risk for Fabry disease. The chatbot would then break down the key factors contributing to this classification, such as specific cardiac abnormalities or genetic mutations, and explain how these factors influence the model's prediction. This conversational approach helps demystify the model's outputs and provides a user-friendly means of engaging with complex data patterns.

8.3.5 Aligning Model Outputs with Clinical Expectations and Usability Through Iterative Testing and Feedback

Through close collaboration and regular engagements with clinicians throughout the tool development cycle, we refined the model and its interface to align with clinical objectives and enhance usability. Regular testing sessions and feedback loops helped identify areas for improvement in the model's performance, the clarity of its explanations, and the overall user experience. This iterative approach ensures that the final product is accurate but also practical and trustworthy in the eyes of healthcare professionals.

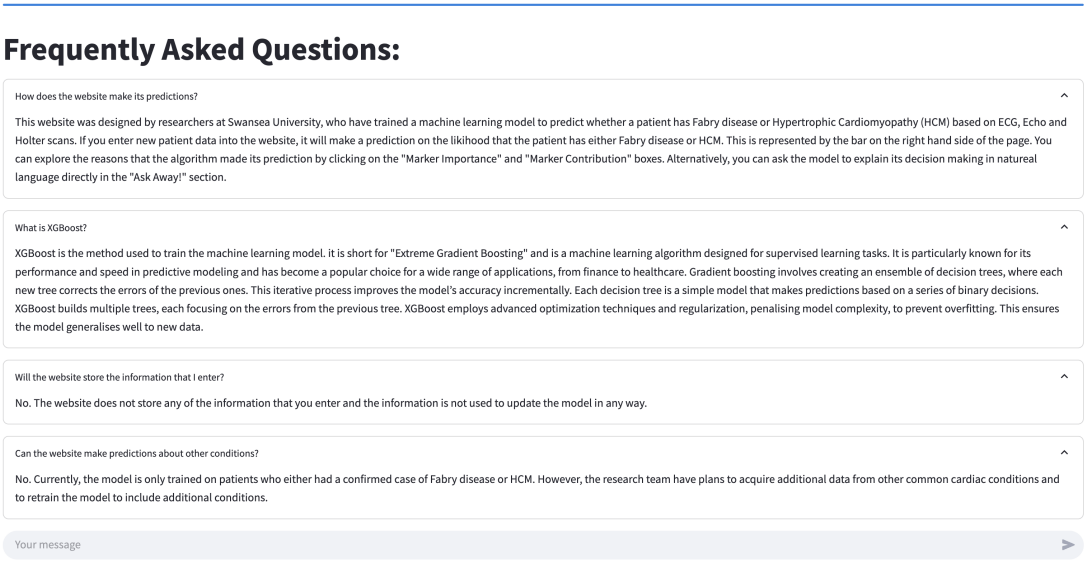


Figure 8.3: The chatbot interface for Fabry disease diagnosis, displaying frequently asked questions (FAQs) and the interactive feature that allows clinicians to explore the model's decision-making process.

In summary, enhancing clinicians' understanding of model predictions involves a multi-faceted approach that provides clear visualisations enabling seamless data input and scenario exploration, utilising conversational LLM chatbots for explanations, and aligning model outputs with clinical expectations through iterative testing and feedback. By implementing these strategies, we aim to create a diagnostic tool that is transparent, interpretable, and truly useful for clinicians in the context of Fabry disease diagnosis and management.

8.4 Integrating ML and LLMs

8.4.1 Incorporating a LangChain-Based LLM Chatbot for Interpreting Model Decision-Making Processes

To enhance the interpretability of the ML model's decision-making processes, we integrated a LangChain-based LLMs chatbot into the clinical risk calculator interface. This chatbot is an interactive tool for clinicians to explore the reasoning behind the model's predictions. It enables clinicians to engage in a dialogue that provides clear, natural language explanations of how the model arrived at its conclusions.

Through interactions with the chatbot, clinicians can probe the underlying factors influencing risk calculations. This helps them understand the model's considerations and limitations,

such as which patient characteristics (e.g., age, ECG findings) played the most significant role in the model's output. The LangChain framework facilitates the generation of context-specific, intuitive explanations, allowing clinicians to navigate complex medical data with greater clarity. By offering this natural language feedback, the LLM makes the decision-making process more accessible, fostering clinician trust and comprehension.

8.4.2 Fine-Tuning the Model and Chatbot for Enhanced Decision Pathway Interpretation

The ML model was iteratively trained on high-quality, diverse datasets to account for the heterogeneous nature of FD. This process included labelled patient data, such as clinical test results, historical diagnoses, and longitudinal records, ensuring the model captured the complexities of FDs clinical presentation. At the same time, the LLM chatbot's responses were fine-tuned to align with clinical reasoning, ensuring that the language generated by the model matched medical terminology and was understandable in a clinical context.

The fine-tuning process was collaborative, involving continuous testing and feedback from clinicians. This iterative approach allowed for the refinement of the model's predictive accuracy and the chatbot's ability to generate relevant, actionable explanations. The model and chatbot were tailored to meet the specific needs of clinicians, making them both useful and reliable tools in real-world clinical settings. This fine-tuning process ensured that the LLM provided consistent, clinically appropriate explanations that clinicians would trust.

8.4.3 Validating the Model and Applying Explainability Techniques

We conducted rigorous validation tests using real-world clinical data and established benchmarks to ensure that the FD-HCM ML model performs reliably in clinical settings. These validation steps are vital for building clinician confidence in the model's predictions and supporting broader adoption.

In addition to performance validation, we employed explainability techniques such as SHAP values and LIME to enhance transparency. SHAP values, for example, quantify how each feature (e.g., ECG findings, patient age) contributes to the model's predictions. This allows clinicians to trace the most influential factors driving the model's diagnostic suggestion. Counterfactual explanations were also implemented to show how a change in input features (e.g., if a patient's age differed) would alter the model's output.

These explainability methods ensure clinicians can trust the model's decisions by making the underlying reasoning visible. This transparency is essential in clinical practice, where healthcare professionals must understand the model's predictions and their rationale.

8.4.4 Addressing Data Limitations and Biases

Addressing challenges such as data limitations and biases is essential in developing clinically interpretable systems. High-quality, diverse datasets are critical for training robust AI systems; however, they are often scarce, especially in rare diseases like FD, due to privacy concerns, logistical hurdles, and the limited number of affected patients. Moreover, biases in training data—such as underrepresenting certain demographic groups—can lead to skewed model outputs, exacerbating disparities in healthcare.

To mitigate these issues, our approach includes sourcing anonymised data from diverse patient populations and implementing feedback mechanisms for continuous model refinement. We also address potential biases by ensuring the data is representative and conducting regular audits. Including expert feedback from anonymised clinical scenarios helps refine the chatbot's decision-making and ensures that its responses are unbiased and applicable to diverse patient populations.

8.5 User Studies of the Prototype

To evaluate the effectiveness and usability of our AI-powered diagnostic tool for Fabry disease, we conducted a focus group study with cardiologists at The University Hospital of Wales (UHW), Cardiff. The study aimed to gather valuable feedback on the tool's features, explainability, and potential role in the clinical workflow.

8.5.1 Study Design and Participants

We conducted a semi-structured group discussion with six cardiologists, including three consultants and three junior doctors. The author facilitated the conversation using a Mentimeter presentation with videos and photos of the prototype. Participants were encouraged to share their views verbally and could post their opinions anonymously on a shared platform.

8.5.2 Ethical Approval and Participant Consent

Ethical approval for this study was obtained from the relevant institutional review board at the University Hospital of Wales (UHW), Cardiff. As the study was retrospective and based on anonymised patient data previously stored within the NHS system, individual consent was not required. Consent for using medical data in research had already been sought as part of standard clinical practice. All data used in the study were anonymised to ensure confidentiality and compliance with data protection regulations.

8.5.3 Feedback on Tool Features

The strongest feedback from participants indicated that manual data input was too time-consuming, which would prevent them from using the tool, as shown in Figure 8.4. The consultants reported that they would like to know the probability of a patient having Fabry disease before seeing the interface, particularly in unexplained left ventricular hypertrophy cases. They also felt that forecasting features to detect early cardiac involvement would be a valuable addition, as this would influence the patient's management.

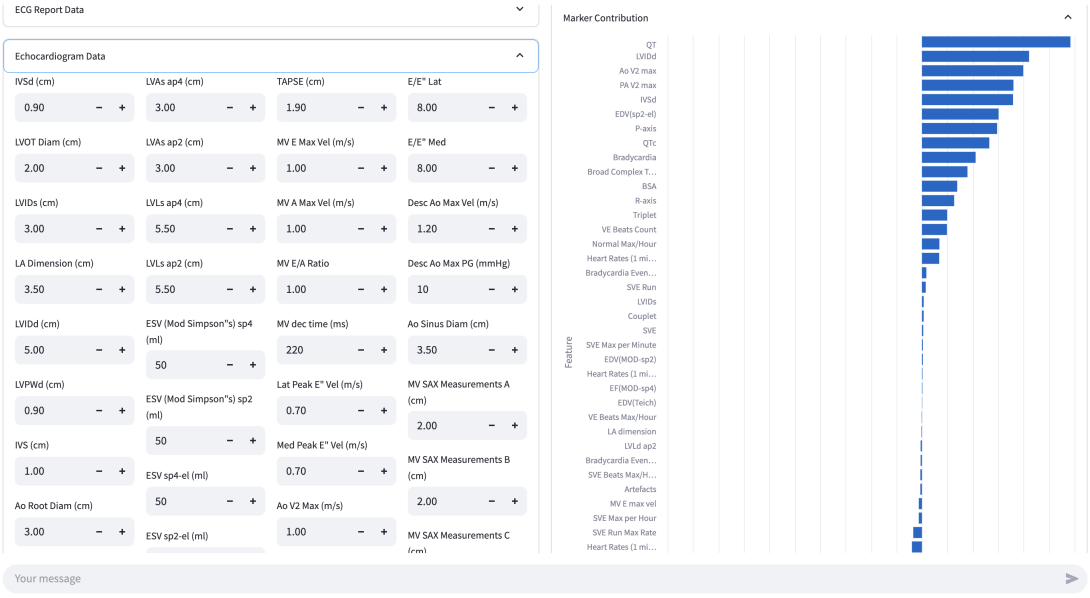


Figure 8.4: User interface of the Fabry disease diagnostic tool, showing input fields for patient data and the model's prediction output alongside the interactive chatbot for exploring the decision-making process.

8.5.4 Feedback on Explainability

The consultants indicated they would not use the chatbot feature to interrogate the data, as they found it relied too heavily on stock 'template' responses, which diminished its trustworthiness. The junior doctors reported that the chatbot language and the feature importance graphs were clear. Still, they suggested that the SHAP values graph be renamed, as this term is unfamiliar to a clinical audience.

8.5.5 Feedback on Workflow Integration

The consultants did not think they would use the tool, as they felt their expertise in identifying suspected FD would not need confirmation from an AI tool, and they would not trust it if it contradicted their judgment. However, the junior doctors reported that this tool would be helpful in a secondary care setting if they were less familiar with the signs of FD, as it could prompt another possible clinical pathway to consider. The consultants also suggested that such a tool could be used in primary care settings to screen routine ECG scans in GP settings. They felt that including information on Fabry disease's other symptom profiles in neurology and renal settings would be helpful.

8.5.6 Implications for Future Development and Research

The feedback from the focus group has important implications for the future development of our AI-powered diagnostic tool. To address the concerns about manual data input, we plan to include a feature that automatically populates input values from clinical reports, significantly reducing the time required to receive a prediction.

While the feature importance table was interpretable by the clinicians and effective in explaining the tool's prediction, the consultants did not see themselves interacting with the language model despite its ability to provide further explanations. This highlights the need for additional research on the role of language models in clinical decision-support tools and how they can be designed better to meet the needs and expectations of experienced clinicians.

The varying opinions on workflow integration between consultants and junior doctors suggest that the tool may be more beneficial for less experienced clinicians in secondary care settings. Few studies have investigated the optimal integration of AI diagnostic tools into existing clinical workflows, and our findings underscore the importance of considering the specific needs and preferences of different user groups.

Although the focus group included an equal number of consultants and junior doctors, the consultants dominated the conversation despite attempts to allow anonymous sharing of opinions. This limitation and the potential for bias in a group discussion format highlight the need for individual interviews to explore explainability and workflow integration in greater depth, particularly with junior doctors who are most likely to benefit from the tool in secondary care settings.

As we continue to develop and validate our AI-powered diagnostic tool, we remain committed to conducting rigorous user studies and incorporating feedback from healthcare professionals to ensure that it meets their needs and ultimately improves patient outcomes in managing Fabry disease.

8.6 Continuous Improvement, Limitations, and Future Directions

8.6.1 Incorporating User Feedback for Long-Term Effectiveness and Relevance

The focus group study conducted with cardiologists at The University Hospital of Wales (UHW), Cardiff, provided valuable insights into the strengths and limitations of our AI-powered diagnostic tool for Fabry disease. To ensure the long-term effectiveness and relevance of the tool, we are committed to incorporating this feedback into future iterations.

A key area for improvement identified was the need for an automatic population of input values from clinical reports, as manual data entry was seen as too time-consuming. We plan to address this issue by developing a feature to extract relevant data from EHRs and automatically populate the tool's input fields. This will reduce the time clinicians need to receive a prediction and streamline the tool's use in clinical practice.

Additionally, user feedback revealed varying opinions on the chatbot feature and the interpretability of the SHAP values graph. While junior doctors found the chatbot language and feature importance graphs clear, consultants were less inclined to interact with the language model. This highlights the need for further research on designing language models that better meet the expectations of experienced clinicians. We plan to conduct additional user studies, particularly with junior doctors, to explore the role of language models in clinical decision support tools and optimise them for different user groups.

8.6.2 Exploring Novel Approaches for Enhanced Explainability and Workflow Integration

The feedback from user studies further underscores the importance of explainability and workflow integration when adopting AI-powered diagnostic tools. Although the feature importance table was effective for clinicians in explaining the tool's predictions, there is room for improvement in how the tool communicates its reasoning.

One potential avenue for enhancing explainability is embedding counterfactual explanations, which show how a model's prediction would change if certain input features were different. This approach could help clinicians better understand the model's sensitivity to patient-specific characteristics and provide insights into areas for further investigation.

Another area for exploration is integrating the tool into existing clinical workflows. The focus group study revealed differing opinions between consultants and junior doctors regarding how the tool could be used in practice. Junior doctors saw more potential for the tool in secondary care settings, while consultants were more cautious. To address these differences, we plan to conduct additional user studies with a larger and more diverse sample of healthcare professionals, including those in primary care and other specialities such as neurology and nephrology. These studies will help us better understand the specific needs and preferences of different user groups and inform the development of a tool that can seamlessly integrate into various clinical settings.

8.6.3 Extending the Diagnostic Tool to Other Rare Diseases and Clinical Settings

While the current focus is on improving the diagnosis of Fabry disease, we recognise the potential for extending our AI-powered diagnostic tool to other rare diseases and clinical settings. Feedback from the focus group study suggests that the tool could help screen routine ECG scans in primary care settings and identify other symptom profiles of Fabry disease in neurology and renal settings.

To extend the tool to other rare diseases, we plan to collaborate with experts in those areas to curate high-quality, diverse datasets and develop models that can accurately identify the specific diagnostic criteria and risk factors associated with each condition. This will require careful consideration of the unique challenges and opportunities presented by different rare diseases and the particular needs and preferences of the healthcare professionals who manage them.

By extending the tool to a broader range of rare diseases and clinical settings, we aim to significantly impact the early detection and management of these conditions, ultimately improving patient outcomes and quality of life.

8.6.4 Addressing Ethical Considerations and Potential Biases in AI-Powered Diagnostic Tools

As we continue refining our AI-powered diagnostic tool, we are committed to addressing the ethical considerations and potential biases that may arise in using such technologies. The focus group study highlighted the importance of trust in adopting AI tools, with consultants expressing reluctance to use a tool that contradicted their judgment.

To build trust in our tool, we will conduct rigorous validation studies to assess its performance across different patient populations and clinical settings. We will also engage in ongoing dialogue with healthcare professionals and patients to better understand their concerns and expectations surrounding the use of AI in healthcare.

Another critical ethical consideration is the potential for bias in the training data and algorithms used to develop our tool. To mitigate this risk, we will regularly audit our datasets and models for potential sources of bias, ensuring they represent the diverse patient populations we aim to serve. We will collaborate with bioethics and health equity experts to establish best practices for the ethical development and deployment of AI-powered diagnostic tools.

By proactively addressing these ethical considerations and potential biases, we aim to build trust in our tool and ensure it contributes to fair and equitable healthcare for all patients.

8.6.5 Limitations of the Study

While this study has contributed to developing an AI-powered diagnostic tool for Fabry disease, several limitations should be considered. These limitations inform future development and offer avenues for improvement.

One limitation is the small sample size of the user study, which included only six participants. A larger, more diverse group of clinicians from various specialities and experience levels would improve the robustness of the findings and provide a more comprehensive understanding of the prototype's functionality across different clinical settings.

Due to logistical constraints, participants interacted with the prototype only via videos and photos, rather than hands-on usage. This limited the ability to assess its usability and real-

time performance. Allowing clinicians to interact with the tool directly would offer a better understanding of how it integrates into clinical workflows and aids decision-making.

Furthermore, detailed demographic data about participants, such as their prior experience with AI tools, was not collected, which limits the ability to analyse feedback based on different levels of comfort with technology. Including direct quotes from participants could also have strengthened the real-world perspective of the feedback.

The study did not include a comparative evaluation of the prototype against traditional diagnostic methods, which makes it difficult to ascertain whether the tool improves diagnostic accuracy or efficiency. Future work should include empirical evidence from clinical settings that directly compares the tool with conventional methods.

While explainability features, such as SHAP values and the LLM-driven chatbot, were integrated, more empirical evidence is required to demonstrate how these features build clinician trust and adoption. A comprehensive evaluation, involving diverse clinicians using the prototype in real clinical scenarios, is necessary to assess the tool's effectiveness in improving decision-making.

These limitations offer several opportunities for further research, including larger and more diverse user studies, real-world hands-on evaluations, and comparative studies against traditional diagnostic tools. Addressing these limitations will enhance the diagnostic tool's clinical relevance, trustworthiness, and effectiveness.

8.7 Summary

We developed an AI-powered diagnostic tool that integrates ML and LLM to enhance clinical decision-making in diagnosing FD. The tool uses ML techniques such as SHAP values and LIME for feature importance and local explanations, alongside partial dependence plots to visualise feature effects. Including a LangChain-based LLM chatbot provides conversational explanations that highlight the most influential markers and offer clinical context, making the model's decision-making process more accessible to clinicians.

User studies showed that the tool effectively supported clinicians in understanding key diagnostic markers, fostering greater confidence in clinical decisions. Clinicians found the feature importance visualisations, particularly the SHAP values, helpful in understanding how specific patient data influenced the model's output. The tool also provided valuable insights by explaining the decision-making process in a clear, understandable way, which allowed clinicians to apply this information in real-world scenarios.

This approach emphasises the importance of explainability and interpretability in AI-powered medical tools, enabling clinicians to trust and integrate the model's predictions into their clinical judgment. By highlighting influential markers and offering a transparent rationale, the tool supports more informed clinical decision-making, ultimately improving patient care.

The tool represents a significant step forward in AI-assisted clinical decision support. We plan to continue refining it, ensuring it aligns with clinical workflows and addresses challenges such as bias and data privacy. Ongoing collaboration with healthcare providers will help optimise the tool's integration into practice, ensuring it provides valuable clinical support for rare disease diagnosis and management.

Chapter 9

Conclusion and Future Work

Contents

Our work has explored the challenges and opportunities of modelling high-fidelity clinical data using interpretable ML techniques for tackling imbalance in healthcare datasets. Motivated by the pressing need to improve diagnostic processes for rare diseases and atypical presentations, which often lead to prolonged patient journeys and sub-optimal outcomes, our research has focused on developing novel methodologies to enhance early detection and clinical decision support.

Through interdisciplinary studies, we have addressed critical gaps in the current state-of-the-art. These include handling data imbalance and missingness in clinical datasets, incorporating temporal patterns and structured medical knowledge into predictive models, and integrating large language models to augment interpretability. By applying these innovative approaches to real-world clinical scenarios, such as the differentiation of Fabry disease from hypertrophic cardiomyopathy, we have demonstrated the potential of ML to transform rare disease diagnosis and management.

The insights and advancements presented in this thesis hold significant implications for clinical practice, patient outcomes, and the efficiency of healthcare systems. By enabling earlier and more accurate identification of rare conditions, our work contributes to developing a more proactive, personalised, and data-driven approach to patient care, ultimately aiming

to alleviate the burden of diagnostic odysseys and improve the quality of life for individuals affected by rare diseases.

In this concluding chapter, we will first summarise the key contributions of our research, highlighting the methodological innovations and their impact on the field. We will then reflect on our work's broader implications and future potential, discussing how it aligns with the ongoing digital transformation of healthcare and the shift towards value-based, patient-centred care. Connecting back to the motivating patient stories introduced at the outset of this thesis, we will underscore the real-world significance of our findings and the urgent need for continued advancements in rare disease diagnosis. Finally, we will outline promising avenues for future research, building upon the foundation laid by this thesis to drive further progress in the field.

9.1 Contributions

The key contributions of this work can be summarised as follows:

Understanding the Patient Journey in the Digital Age

We offered a comprehensive overview of how digital transformation reshapes the patient journey within the UK healthcare system. By analysing the roles of primary, secondary, and tertiary care, we highlighted the complexities and challenges in diagnosing rare diseases and atypical presentations, emphasising the need for data-driven approaches to improve patient outcomes.

Addressing Data Imbalance and Missingness in Clinical Settings

Recognising that clinical datasets often suffer from imbalance and missing data, we designed a clinically aligned approach to modelling ICU data. By implementing sophisticated data preprocessing techniques and feature engineering, we improved the quality of the datasets, enabling more accurate and reliable ML models for early sepsis detection.

Enhancing Rare Event Detection through ML

We developed novel ML models tailored for rare event detection, particularly focusing on sepsis and COVID-19 complications. By incorporating temporal patterns and self-attention mechanisms, our models demonstrated improved predictive performance, enabling earlier and more accurate identification of at-risk patients.

Integrating Structured Medical Knowledge into Language Models

To improve the classification of rare diseases and enhance interpretability, we proposed

a method for infusing ontological structures into language models. This approach leveraged medical ontologies to augment the representation of clinical terms, resulting in a more nuanced understanding and better performance in rare disease detection tasks.

Differentiating Complex Cardiac Diseases with Similar Presentations

Focusing on FD and its similarities with hypertrophic cardiomyopathy, we applied ML techniques to distinguish between these conditions effectively. Our models utilised high-fidelity ECG and echocardiogram data, achieving significant accuracy and demonstrating the potential for assisting clinicians in making more informed diagnostic decisions.

Augmenting ML Predictions with Large Language Models (LLMs)

We explored the integration of LLM to enhance the interpretability of ML predictions for clinicians. By providing natural language explanations alongside risk assessments, we improved the usability of predictive models in clinical settings, fostering greater trust and facilitating their adoption in practice.

9.2 Broader Implications and Future Impact

The novel methodologies developed in this thesis for detecting rare diseases and atypical presentations using interpretable ML can potentially transform diagnostic processes and improve patient outcomes on a larger scale. These approaches could significantly reduce the economic burden on healthcare systems by enabling earlier and more accurate diagnoses.

Additionally, the techniques presented here for capturing the heterogeneity of rare diseases at the individual patient level represent a crucial step towards advancing personalised medicine. By accounting for the unique characteristics and progression patterns of rare conditions in each patient, clinicians can tailor interventions and management strategies to optimise outcomes.

The continued refinement and expansion of these ML models to incorporate additional data modalities, including genomics and imaging, could further enhance their predictive power and clinical utility. Integrating these tools into EHR systems and clinical decision support platforms would facilitate their adoption into routine practice, extending their impact to larger patient populations.

Ultimately, the knowledge and insights gained from this work lay the foundation for a paradigm shift in rare disease diagnosis from a reactive, fragmented process to a proactive, data-driven approach that empowers clinicians to provide timely, targeted care to patients who have historically faced significant challenges in receiving accurate diagnoses and appropriate

treatments. By bridging the gap between cutting-edge ML research and real-world clinical applications, our work paves the way for meaningful advancements in the field of rare diseases.

9.3 Connecting Back to Patient Stories

The stories of patients like Emily, who have navigated the complex and often frustrating journey of obtaining a rare disease diagnosis, were the driving force behind our work. For individuals living with conditions like Fabry disease, delays in diagnosis can lead to irreversible organ damage and significantly reduced quality of life.

The ML models and interpretability techniques developed here aim to address these challenges head-on by equipping clinicians with powerful tools to recognise the subtle signs and atypical presentations of rare diseases earlier in the diagnostic process. Had such tools been available during Emily's diagnostic odysseys, her paths to receiving an accurate diagnosis and appropriate care may have been significantly shortened, potentially mitigating the impact of the disease on their life.

While the road ahead remains long, the advancements presented throughout our work represent tangible steps towards a future in which stories like Emily's become the exception rather than the norm. By harnessing the power of data and ML, while keeping the human element at the forefront, we can work towards a healthcare system better equipped to identify and support patients with rare diseases, ensuring they receive the timely, personalised care they need and deserve.

In this way, the technical innovations detailed in this work are inextricably linked to the lived experiences of patients, serving as a poignant reminder of the real-world impact that advancements in rare disease diagnosis can have on individuals, families, and communities. As we look to the future, this connection between cutting-edge research and tangible improvements in patient care will continue to drive progress in the field, bringing us closer to a world where no patient's story is defined by the challenges of navigating a diagnostic odyssey.

9.4 The Path Forward

As we look towards the future, it is clear that the groundwork laid by this thesis has opened up many exciting possibilities for advancing the identification and management of rare phenomena in healthcare through applying interpretable ML techniques. The potential impact of

this research is far-reaching, promising to revolutionise how we approach rare diseases and ultimately improve patient outcomes.

A key area for future exploration is expanding our models to incorporate multi-modal data sources. By integrating diverse data modalities such as genetic information, imaging data, and patient-reported outcomes, we can paint a more comprehensive picture of patient health and enhance the performance and generalisability of our predictive models. However, this integration process is not without its challenges. Ensuring data preprocessing and standardisation following clinical guidelines is crucial to maintaining the highest quality and integrity standards. This involves meticulous data cleaning, harmonisation, and validation to guarantee the reliability and accuracy of the integrated datasets, which form the bedrock of our modelling efforts.

Another possible advancement of our model is the real-time implementation and evaluation of our model in clinical settings. Collaborating closely with healthcare providers to pilot these models in real-world environments will provide invaluable insights into their practical utility and highlight areas for refinement and optimisation. Investing in research initiatives that explore novel methodologies, validate existing models, and rigorously assess their impact on patient outcomes is essential to support this endeavour. This includes securing funding for large-scale studies, pilot projects, and implementation research to comprehensively evaluate the effectiveness and feasibility of our approaches across diverse healthcare contexts.

While leveraging patient data, we must remain vigilant in addressing ethical and privacy concerns. Future research must prioritise exploring secure data handling methods, such as federated learning and differential privacy techniques, to safeguard patient information while enabling the powerful insights that drive our progress. Building a robust infrastructure for secure data sharing, storage, and analysis is a fundamental prerequisite for facilitating collaboration and accelerating advancements in this field.

Central to the success of using ML tools in clinical settings is placing greater emphasis on the explainability and interpretability of the models. By creating models that provide clear, intuitive explanations for their predictions and align seamlessly with existing clinical knowledge, we can foster a deeper understanding and acceptance among healthcare professionals. Techniques such as feature importance analysis, rule-based models, and visual explanations hold great promise to enhance the interpretability of clinical models, making them more actionable and valuable for clinicians in their daily practice.

Longitudinal studies and exploring temporal dynamics represent another frontier for future

research. By investigating the evolution of disease progression and patient health over extended periods, we can uncover more profound insights into the complex interplay of factors that shape outcomes. Longitudinal studies enable our models to account for changes over time, potentially unlocking new opportunities for early detection and intervention. This approach resonates strongly with the principles of VBHC, which places patient outcomes at the centre of all decision-making. As future models are being developed, we must incorporate patient-centric metrics into our models and evaluate their contributions to improved quality of life and care satisfaction.

Lastly, underpinning all these efforts is the importance of fostering interdisciplinary collaboration. The challenges of rare event detection and management are complex and multifaceted, requiring the collective expertise of data scientists, clinicians, and domain experts. By bringing together diverse perspectives and skill sets, future researchers could ensure that the models developed are technically robust, clinically relevant, and actionable. Regular workshops, joint research projects, and knowledge-sharing platforms are powerful catalysts for interdisciplinary collaboration, promoting a shared understanding of the goals and challenges all stakeholders face.

In conclusion, the future of rare phenomenon identification in healthcare is brimming with potential, and this thesis has laid the foundation for a new era of innovation and progress. By building upon the groundwork laid by this thesis, collaborating with interdisciplinary teams, and relentlessly pursuing innovation, we can create a future where the early and accurate diagnosis of rare conditions is not the exception but the norm. Together, we can harness the power of interpretable ML to illuminate the darkest corners of rare diseases, bringing hope and healing to countless lives and shaping a brighter, healthier future for all.

Bibliography

[1] NHS England, "The healthcare ecosystem," https://digital.nhs.uk/developer/guides-and-documentation/introduction-to-healthcare-technology/the-healthcare-ecosystem, 2023, accessed: 2025-05-11. [Online]. Available: https://digital.nhs.uk/developer/guides-and-documentation/introduction-to-healthcare-technology/the-healthcare-ecosystem

[2] G. E. Programme, "Introduction to rare disease," https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/introduction-to-rare-disease/, 2024, accessed: 2025-05-19. [Online]. Available: https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/introduction-to-rare-disease/

[3] A. A. Mitani and S. Haneuse, "Small data challenges of studying rare diseases," *JAMA Network Open*, vol. 3, no. 3, p. e201965, 2020.

[4] P. Z. Zimmet, D. J. Magliano, W. H. Herman, and J. E. Shaw, "Diabetes: a 21st century challenge," *The lancet Diabetes & endocrinology*, vol. 2, no. 1, pp. 56–64, 2014.

[5] F. Hobbs, M. Piepoli, A. Hoes, S. Agewall, C. Albus, C. Brotons, A. Catapano, M. Cooney, U. Corra, B. Cosyns *et al.*, "2016 european guidelines on cardiovascular disease prevention in clinical practice." *European heart journal*, vol. 37, no. 29, 2016.

[6] M. M. Koo, K. Unger-Saldaña, A. D. Mwaka, M. Corbex, O. Ginsburg, F. M. Walter, N. Calanzani, J. Moodley, G. P. Rubin, and G. Lyratzopoulos, "Conceptual framework to guide early diagnosis programs for symptomatic cancer as part of global cancer control," *JCO Global Oncology*, no. 7, pp. 35–45, 2021, pMID: 33405957. [Online]. Available: https://ascopubs.org/doi/abs/10.1200/GO.20.00310

[7] StatsWales, "Population estimates by local authority and age," 2020, accessed: 2021-09-01. [Online]. Avail-

able: https://statswales.gov.wales/Catalogue/Population-and-Migration/Population/Estimates/Local-Authority/populationestimates-by-localauthority-age

[8] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.

[9] Roche, "Heart failure:the hidden costs of late diagnosis," 2020, accessed: 2025-02-01. [Online]. Available: https://gov.wales/healthier-wales-long-term-plan-health-and-social-care

[10] T. P. Hanna, W. D. King, S. Thibodeau, M. Jalink, G. A. Paulin, E. Harvey-Jones, D. E. O'Sullivan, C. M. Booth, R. Sullivan, and A. Aggarwal, "Mortality due to cancer treatment delay: systematic review and meta-analysis," *bmj*, vol. 371, 2020.

[11] G. F. Pilz, F. Weber, W. G. Mueller, and J. R. Schaefer, "Statistical methods to support diagnosing rare diseases," *ResearchSquare*, 2020. [Online]. Available: https://doi.org/10.21203/rs.2.23479/v1

[12] G. E. Programme, "The diagnostic odyssey in rare disease," 2024, accessed: 2024-05-10. [Online]. Available: https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/the-diagnostic-odyssey-in-rare-disease/

[13] E. Hay, F. Elmslie, P. Lanyon, and T. Cole, "The diagnostic odyssey in rare diseases; a task and finish group report for the department of health and social care," *NIHR Open Res*, vol. 2, no. 3, p. 3, 2022.

[14] M. E. Porter and T. H. Lee, "The strategy that will fix health care," *Harvard Business Review*, vol. 99, no. 6, pp. 42–65, 2021. [Online]. Available: https://hbr.org/2021/06/the-strategy-that-will-fix-health-care

[15] W. Government, "A healthier wales: our plan for health and social care," 2022, accessed: 2024-09-01. [Online]. Available: https://gov.wales/healthier-wales-long-term-plan-health-and-social-care

[16] Welsh Government, "The rare disease implementation plan for wales," 2024, accessed: 2024-09-01. [Online]. Available: https://executive.nhs.wales/functions/networks-and-planning/rare-diseases/rdig-documents/action-plan-refresh-jan-2024/

[17] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Med. Res. Methodol.*, vol. 19, no. 1, p. 64, Mar. 2019.

[18] EY, "The value of healthcare data," EY Global, 2023, accessed: 2023-05-26. [Online]. Available: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/life-sciences/life-sciences-pdfs/ey-value-of-health-care-data-v20-final.pdf

[19] W. H. Organization, "Building resilient health systems: Lessons from the covid-19 pandemic," World Health Organization, Tech. Rep., 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240057954

[20] NHS, "About the nhs," 2024. [Online]. Available: https://www.stepintothenhs.nhs.uk/about-the-nhs

[21] W. Government, "A plan for digital health and social care: policy paper," June 2022. [Online]. Available: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care

[22] L. Edwards, J. Pickett, D. M. Ashcroft, H. Dambha-Miller, A. Majeed, C. Mallen, I. Petersen, N. Qureshi, T. van Staa, G. Abel, C. Carvalho, R. Denholm, E. Kontopantelis, A. Macaulay, and J. Macleod, "UK research data resources based on primary care electronic health records: review and summary for potential users," *BJGP Open*, vol. 7, no. 3, p. BJGPO.2023.0057, Sep. 2023.

[23] N. Garcelon, A. Burgun, R. Salomon, and A. Neuraz, "Electronic health records for the diagnosis of rare diseases," *Kidney Int.*, vol. 97, no. 4, pp. 676–686, Apr. 2020.

[24] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS One*, vol. 7, no. 1, p. e30412, Jan. 2012.

[25] A. Moroni, L. Tondi, V. Milani, M. Pieroni, F. Pieruzzi, F. Bevilacqua, G. Pasqualin, K. Chow, S. Pica, M. Lombardi *et al.*, "Left atrial remodeling in hypertrophic cardiomyopathy and fabry disease: A cmr-based head-to-head comparison and outcome analysis," *International Journal of Cardiology*, vol. 393, p. 131357, 2023.

[26] NHS, "The healthcare ecosystem," Unknown. [Online]. Available: https://digital.nhs.uk/developer/guides-and-documentation/introduction-to-healthcare-technology/the-healthcare-ecosystem

[27] J. Sedlakova, P. Daniore, A. Horn Wintsch, M. Wolf, M. Stanikic, C. Haag, C. Sieber, G. Schneider, K. Staub, D. Alois Ettlin, O. Grübner, F. Rinaldi, V. von Wyl, and University of Zurich Digital Society Initiative (UZH-DSI) Health Community, "Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review," *PLOS Digit. Health*, vol. 2, no. 10, p. e0000347, Oct. 2023.

[28] J. S. Tulloch, M. B. Beadsworth, R. Vivancos, A. D. Radford, J. C. Warner, and R. M. Christley, "GP coding behaviour for non-specific clinical presentations: a pilot study," *BJGP Open*, vol. 4, no. 3, p. bjgpopen20X101050, Aug. 2020.

[29] E. Herrett, A. D. Shah, R. Boggon, S. Denaxas, L. Smeeth, T. van Staa, A. Timmis, and H. Hemingway, "Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study," *Bmj*, vol. 346, 2013.

[30] C. Li, D. L. Mowery, X. Ma, R. Yang, U. Vurgun, S. Hwang, H. K. Donnelly, H. Bandhey, Z. Akhtar, Y. Senathirajah, E. M. Sadhu, E. Getzen, P. J. Freda, Q. Long, and M. J. Becich, "Realizing the potential of social determinants data: A scoping review of approaches for screening, linkage, extraction, analysis and interventions," Feb. 2024.

[31] J. Rigg, H. Lodhi, and P. Nasuti, "Using machine learning to detect patients with undiagnosed rare diseases: an application of support vector machines to a rare oncology disease," *Value in Health*, vol. 18, no. 7, p. A705, 2015.

[32] N. N. Taleb, *The black swan: The impact of the highly improbable*. Random House, 2007, vol. 2.

[33] T. Willmen, L. Willmen, A. Pankow, S. Ronicke, H. Gabriel, and A. D. Wagner, "Rare diseases: why is a rapid referral to an expert center so important?" *BMC Health Services Research*, vol. 23, no. 1, p. 904, 2023.

[34] C. Feng, L. Li, and C. Xu, "Advancements in predicting and modeling rare event outcomes for enhanced decision-making," *BMC Med. Res. Methodol.*, vol. 23, no. 1, p. 243, Oct. 2023.

[35] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, "Comparison of evaluation metrics in classification applications with imbalanced datasets," in *2008 seventh international conference on machine learning and applications*. IEEE, 2008, pp. 777–782.

[36] C. Shyalika, R. Wickramarachchi, and A. Sheth, "A comprehensive survey on rare event prediction," 2024. [Online]. Available: https://arxiv.org/abs/2309.11356

[37] B. Malkiel, "The black monday stock market crash," *The Journal of Portfolio Management*, vol. 15, no. 3, pp. 35–42, 1989.

[38] M. Sullivan, "Insurance and catastrophic natural disasters," *Journal of Risk and Insurance*, vol. 78, no. 3, pp. 613–640, 2011.

[39] R. Gallo, "The hiv-aids epidemic and the need for global health equity," *Lancet*, vol. 358, no. 9285, pp. 233–239, 2001.

[40] M. Feldman, "Challenges in diagnosing rare diseases: A review," *Journal of Clinical Medicine*, vol. 6, no. 4, pp. 42–50, 2017.

[41] S. Rodríguez-Martín, E. Martín-Merino, V. Lerma, A. Rodríguez-Miguel, O. González, C. González-Herrada, E. Ramírez, T. Bellón, and F. J. de Abajo, "Active surveillance of severe cutaneous adverse reactions: A case-population approach using a registry and a health care database," *Pharmacoepidemiology and Drug Safety*, vol. 27, pp. 1042 – 1050, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:51724824

[42] R. C. Hennekam, "Hutchinson–gilford progeria syndrome: review of the phenotype," *American journal of medical genetics Part A*, vol. 140, no. 23, pp. 2603–2624, 2006.

[43] C. L. Ventola, "Big data and pharmacovigilance: data mining for adverse drug events and interactions," *Pharmacy and therapeutics*, vol. 43, no. 6, p. 340, 2018.

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[45] F. Bouchet, J. Rolland, and J. Wouters, "Rare event sampling methods," *Chaos: An interdisciplinary journal of nonlinear science*, vol. 29, no. 8, 2019.

[46] K. McCarthy, B. Zabar, and G. Weiss, "Does cost-sensitive learning beat sampling for classifying rare classes?" in *Proceedings of the 1st international workshop on Utility-based data mining*, 2005, pp. 69–77.

[47] I. T. for the Revision of the International Criteria for Behçet's Disease (ITR-ICBD), F. Davatchi, S. Assaad-Khalil, K. Calamia, J. Crook, B. Sadeghi-Abdollahi, M. Schirmer, T. Tzellos, C. Zouboulis, M. Akhlagi *et al.*, "The international criteria for behçet's disease (icbd): a collaborative study of 27 countries on the sensitivity and specificity of the new criteria," *Journal of the european Academy of Dermatology and Venereology*, vol. 28, no. 3, pp. 338–347, 2014.

[48] A. H. Rowley and S. T. Shulman, "The epidemiology and pathogenesis of kawasaki disease," *Frontiers in pediatrics*, vol. 6, p. 374, 2018.

[49] B. Nikolay, H. Salje, M. J. Hossain, A. D. Khan, H. M. Sazzad, M. Rahman, P. Daszak, U. Ströher, J. R. Pulliam, A. M. Kilpatrick, S. T. Nichol, J. D. Klena, S. Sultana, S. Afroj, S. P. Luby, S. Cauchemez, and E. S. Gurley, "Transmission of nipah virus — 14 years of investigations in bangladesh," *New England Journal of Medicine*, vol. 380, no. 19, pp. 1804–1814, 2019. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1805376

[50] X. Lei and C. A. MacKenzie, "Comparing different models to forecast the number of mass shootings in the united states: An application of forecasting rare event time series data," *PLOS ONE*, vol. 18, no. 6, pp. 1–23, 06 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0287427

[51] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber," in *International conference on machine learning*, vol. 34. sn, 2017, pp. 1–5.

[52] D. M. Jordan, H. M. T. Vy, and R. Do, "A deep learning transformer model predicts high rates of undiagnosed rare disease in large electronic health systems," *medRxiv*, 2023. [Online]. Available: https://www.medrxiv.org/content/early/2023/12/24/2023.12.21.23300393

[53] A. B. Popejoy and S. M. Fullerton, "Genomics is failing on diversity," *Nature*, vol. 538, no. 7624, pp. 161–164, 2016.

[54] A. R. Bentley, S. Callier, and C. N. Rotimi, "Diversity and inclusion in genomic research: why the uneven progress?" *Journal of community genetics*, vol. 8, pp. 255–266, 2017.

[55] S. Wang, M. Lin, T. Ghosal, Y. Ding, and Y. Peng, "Knowledge graph applications in medical imaging analysis: A scoping review," *Health data science*, vol. 2022, 2022.

[56] D. L. Bodian, E. Klein, R. K. Iyer, W. S. Wong, P. Kothiyal, D. Stauffer, K. C. Huddleston, A. D. Gaither, I. Remsburg, A. Khromykh *et al.*, "Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates," *Genetics in Medicine*, vol. 18, no. 3, pp. 221–230, 2016.

[57] P. Mistry, P. Kishnani, C. Wanner, D. Dong, J. Bender, J. Batista, and J. Foster, "Rare lysosomal disease registries: lessons learned over three decades of real-world evidence," *Orphanet Journal of Rare Diseases*, vol. 17, no. 1, p. 362, 2022.

[58] S. Schoenen, J. Verbeeck, L. Koletzko, I. Brambilla, M. Kuchenbuch, M. Dirani, G. Zimmermann, H. Dette, R.-D. Hilgers, G. Molenberghs *et al.*, "Istore: a project on innovative statistical methodologies to improve rare diseases clinical trials in limited populations," *Orphanet Journal of Rare Diseases*, vol. 19, no. 1, p. 96, 2024.

[59] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[60] L. Liu, L. Liu, L.-C. Liang, Z.-q. Zhu, X. Wan, H.-b. Dai, and Q. Huang, "Impact of preoperative anemia on perioperative outcomes in patients undergoing elective colorectal surgery," *Gastroenterology Research and Practice*, vol. 2018, no. 1, p. 2417028, 2018.

[61] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

[62] S. Rafsunjani, R. S. Safa, A. Al Imran, M. S. Rahim, and D. Nandi, "An empirical comparison of missing value imputation techniques on aps failure prediction," *International Journal of Information Technology and Computer Science*, vol. 2, pp. 21–29, 2019.

[63] M. T. Ashraf, K. Dey, and S. Mishra, "Identification of high-risk roadway segments for wrong-way driving crash using rare event modeling and data augmentation techniques," *Accident Analysis & Prevention*, vol. 181, p. 106933, 2023.

[64] Z. S. Y. Wong, "Statistical classification of drug incidents due to look-alike sound-alike mix-ups," *Health informatics journal*, vol. 22, no. 2, pp. 276–292, 2016.

[65] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Computers in Biology and Medicine*, vol. 128, p. 104129, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520304601

[66] H. Jiang, L. Liu, Y. Wang, H. Ji, X. Ma, J. Wu, Y. Huang, X. Wang, R. Gui, Q. Zhao *et al.*, "Machine learning for the prediction of complications in patients after mitral valve surgery," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 771246, 2021.

[67] W. Mao, L. Wang, and N. Feng, "A new fault diagnosis method of bearings based on structural feature selection," *Electronics*, vol. 8, no. 12, p. 1406, 2019.

[68] C. Hong, Y. Xiong, J. Xia, W. Huang, A. Xia, S. Xu, Y. Chen, Z. Xu, H. Chen, and Z. Zhang, "Lasso-based identification of risk factors and development of a prediction model for sepsis patients," *Therapeutics and Clinical Risk Management*, pp. 47–58, 2024.

[69] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Principal component analysis," *Nature methods*, vol. 14, no. 7, pp. 641–643, 2017.

[70] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[71] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *SIGKDD Explor.*, vol. 6, pp. 40–49, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:16051569

[72] S.-J. Yen and Y.-S. Lee, "Cluster-based sampling approaches to imbalanced data distributions," in *International Conference on Data Warehousing and Knowledge Discovery*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:35036696

[73] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst. Man Cybern.*, vol. 2, pp. 408–421, 1972. [Online]. Available: https://api.semanticscholar.org/CorpusID:6699477

[74] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Conference on Artificial Intelligence in Medicine in Europe*, 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:17219136

[75] I. Mani and I. Zhang, "knn approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, no. 1. ICML, 2003, pp. 1–7.

[76] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97, no. 1. Citeseer, 1997, p. 179.

[77] C. Bellinger, C. Drummond, and N. Japkowicz, "Manifold-based synthetic oversampling with manifold conformance estimation," *Machine Learning*, vol. 107, pp. 605–637, 2018.

[78] J. ichiro Fukuchi, "Subsampling and model selection in time series analysis," *Biometrika*, vol. 86, pp. 591–604, 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:120981820

[79] F. Combes, R. Fraiman, and B. Ghattas, "Time series sampling," *Engineering Proceedings*, vol. 18, no. 1, p. 32, 2022.

[80] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[81] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:915058

[82] C. F. Manski, "Anatomy of the selection problem," *Journal of Human Resources*, vol. 24, pp. 343–360, 1989. [Online]. Available: https://api.semanticscholar.org/CorpusID:118631522

[83] Z. I. Botev, A. Ridder, and L. Rojas-Nandayapa, "Semiparametric cross entropy for rare-event simulation," *ERN: Semiparametric & Nonparametric Methods (Topic)*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:40013555

[84] Mansur Arief, Zhiyuan Huang, Guru Koushik Senthil Kumar, Yuanlu Bai, Shengyi He, Wenhao Ding, H. Lam, and Ding Zhao, "Deep Probabilistic Accelerated Evaluation: A Robust Certifiable Rare-Event Simulation Methodology for Black-Box Safety-Critical Systems," *International Conference on Artificial Intelligence and Statistics*, 2021.

[85] D. Straub, I. Papaioannou, and W. Betz, "Bayesian analysis of rare events," *Journal of Computational Physics*, vol. 314, pp. 538–556, 6 2016.

[86] V. Rao, R. Maulik, E. Constantinescu, and M. Anitescu, *A Machine-Learning-Based Importance Sampling Method to Compute Rare Event Probabilities*. Springer International Publishing, 2020, pp. 169–182.

[87] Z. Wang and P. Wang, "Accelerated failure identification sampling for probability analysis of rare events," *Structural and Multidisciplinary Optimization*, vol. 54, no. 1, pp. 137–149, feb 4 2016.

[88] F. Uribe, I. Papaioannou, Y. M. Marzouk, and D. Straub, "Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 9, no. 2, pp. 818–847, 1 2021.

[89] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 2008, pp. 1322–1328.

[90] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:11319376

[91] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.

[92] C. J. Patel, J. Bhattacharya, and A. J. Butte, "An environment-wide association study (ewas) on type 2 diabetes mellitus," *PloS one*, vol. 5, no. 5, p. e10746, 2010.

[93] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone *et al.*, "Passive detection of atrial fibrillation using a commercially available smartwatch," *JAMA cardiology*, vol. 3, no. 5, pp. 409–416, 2018.

[94] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[95] E. J. Cornblath, J. L. Robinson, D. J. Irwin, E. B. Lee, V. M.-Y. Lee, J. Q. Trojanowski, and D. S. Bassett, "Defining and predicting transdiagnostic categories of neurodegenerative disease," *Nature biomedical engineering*, vol. 4, no. 8, pp. 787–800, 2020.

[96] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks: a review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, no. 1, pp. 49–82, 2012. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2011.00714.x

[97] R. C. Ripan, I. H. Sarker, S. M. M. Hossain, M. M. Anwar, R. Nowrozy, M. M. Hoque, and M. H. Furhad, "A data-driven heart disease prediction model through k-means clustering-based anomaly detection," *SN Computer Science*, vol. 2, no. 2, p. 112, 2021.

[98] B. K. Beaulieu-Jones, C. S. Greene *et al.*, "Semi-supervised learning of the electronic health record for phenotype stratification," *Journal of biomedical informatics*, vol. 64, pp. 168–178, 2016.

[99] V. Cheplygina, M. De Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[100] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker *et al.*, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[101] M. Müller, M. Gromicho, M. de Carvalho, and S. C. Madeira, "Explainable models of disease progression in als: Learning from longitudinal clinical data with recurrent neural networks and deep model explanation," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100018, 2021.

[102] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[103] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts, "Deep representation learning of patient data from electronic health records (ehr): A systematic review," *Journal of biomedical informatics*, vol. 115, p. 103671, 2021.

[104] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[105] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[106] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[107] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016. [Online]. Available: https://doi.org/10.1007/s40708-016-0042-6

[108] A. N. Jagannatha and H. Yu, "Bidirectional RNN for medical event detection in electronic health records," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Knight, A. Nenkova, and O. Rambow, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 473–482. [Online]. Available: https://aclanthology.org/N16-1056

[109] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

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

[111] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, pp. 1–13, 2020.

[112] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[113] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632.

[114] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, vol. 10, no. 3. MIT Press, 1999, pp. 61–74.

[115] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 694–699.

[116] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[117] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll'ar, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[118] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[119] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, "A theoretical analysis of ndcg type ranking measures," in *Conference on learning theory*. PMLR, 2013, pp. 25–54.

[120] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[121] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.

[122] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," 2019. [Online]. Available: https://arxiv.org/abs/1905.05134

[123] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, p. 31–57, Jun. 2018. [Online]. Available: https://doi.org/10.1145/3236386.3241340

[124] J. D. Janizek, P. Sturmfels, and S.-I. Lee, "Explaining explanations: Axiomatic feature interactions for deep networks," *Journal of Machine Learning Research*, vol. 22, no. 104, pp. 1–54, 2021.

[125] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[126] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[127] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative ophthalmology & visual science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[128] M. Cecconi, L. Evans, M. Levy, and A. Rhodes, "Sepsis and septic shock," *The Lancet*, vol. 392, no. 10141, pp. 75–87, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673618306962

[129] K. Rudd, S. Johnson, K. Agesa, K. Shackelford, D. Tsoi, D. Kievlan, D. Colombara, K. Ikuta, N. Kissoon, S. Finfer, C. Fleischmann, F. Machado, K. Reinhart, K. Rowan, C. Seymour, S. Watson, E. West, M. D. F. Marinho de Souza, S. Hay, and M. Naghavi, "Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the global burden of disease study," *The Lancet*, vol. 395, pp. 200–211, 01 2020.

[130] F. B. Mayr and D. C. Angus, "Three decades of sepsis in the united kingdom: Tracing the tides of time," *Am. J. Respir. Crit. Care Med.*, vol. 209, no. 5, pp. 468–469, Mar. 2024.

[131] J. D. Forrester, "Sepsis and septic shock," 2024/04 2024. [Online]. Available: https://www.msdmanuals.com/en-gb/professional/critical-care-medicine/sepsis-and-septic-shock/sepsis-and-septic-shock

[132] K. N. Iskander, M. F. Osuchowski, D. J. Stearns-Kurosawa, and et al., "Sepsis: Multiple abnormalities, heterogeneous responses, and evolving understanding," *Physiological Reviews*, vol. 93, pp. 1247–1288, 2013.

[133] S. Lyra, S. Leonhardt, and C. H. Antink, "Early prediction of sepsis using random forest classification for imbalanced clinical data," in *2019 Computing in Cardiology (CinC)*, 2019, pp. 1–4.

[134] B. R. Adegbite, J. R. Edoa, W. F. Ndzebe Ndoumba, L. B. Dimessa Mbadinga, G. Mombo-Ngoma, S. T. Jacob, J. Rylance, T. Hänscheid, A. A. Adegnika, and M. P. Grobusch, "A comparison of different scores for diagnosis and mortality prediction of adults with sepsis in low-and-middle-income countries: a systematic review and meta-analysis," *eClinicalMedicine*, vol. 42, p. 101184, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S258953702100465X

[135] L. M. Castello and F. Gavelli, "Sepsis scoring systems: Mindful use in clinical practice," *Eur. J. Intern. Med.*, vol. 125, pp. 32–35, Jul. 2024.

[136] S. Goodacre, L. Sutton, B. Thomas, O. Hawksworth, K. Iftikhar, S. Croft, G. Fuller, S. Waterhouse, D. Hind, M. Bradburn, M. A. Smyth, G. D. Perkins, M. Millins,

A. Rosser, J. M. Dickson, and M. J. Wilson, "Prehospital early warning scores for adults with suspected sepsis: retrospective diagnostic cohort study," *Emergency Medicine Journal*, vol. 40, no. 11, pp. 768–776, 2023. [Online]. Available: https://emj.bmj.com/content/40/11/768

[137] C. W. Seymour, V. X. Liu, T. J. Iwashyna, and et al., "Assessment of clinical criteria for sepsis," *JAMA*, vol. 315, p. 762, 2016.

[138] L. Zhou, M. Shao, C. Wang, and Y. Wang, "An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context," *Preventive Medicine Reports*, vol. 45, p. 102841, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211335524002560

[139] J. Li, F. Xi, W. Yu, C. Sun, and X. Wang, "Real-time prediction of sepsis in critical trauma patients: Machine learning-based modeling study," *JMIR Form. Res.*, vol. 7, p. e42452, Mar. 2023.

[140] Z. Rayan, M. Alfonse, and A.-B. M. Salem, "Predicting sepsis in the intensive care unit (ICU) through vital signs using support vector machine (SVM)," *Open Bioinforma. J.*, vol. 14, no. 1, pp. 108–113, Nov. 2021.

[141] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review," *Discov. Artif. Intell.*, vol. 4, no. 1, Feb. 2024.

[142] W. H. Organization, *Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions*, 2020.

[143] I. Cortés-Puch and C. S. Hartog, "Opening the debate on the new sepsis definition: Change is not necessarily progress: Revision of the sepsis definition should be based on new scientific insights," *American Journal of Respiratory and Critical Care Medicine*, vol. 194, pp. 16–18, 2016.

[144] T. Desautels, J. Calvert, J. Hoffman, and et al., "Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach," *JMIR Medical Informatics*, vol. 4, p. e28, 2016.

[145] S. P. Shashikumar, M. D. Stanley, I. Sadiq, and et al., "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," *Journal of Electrocardiology*, vol. 50, pp. 739–743, 2017.

[146] S. Nemati, A. Holder, F. Razmi, and et al., "An interpretable machine learning model for accurate prediction of sepsis in the icu," *Critical Care Medicine*, vol. 46, pp. 547–553, 2018.

[147] X. Li, Y. Kang, X. Jia, and et al., "Tasp: A time-phased model for sepsis prediction," in *2019 Computing in Cardiology (CinC)*, 2019, pp. 1–4.

[148] V. Abromavičius, D. Plonis, D. Tarasevičius, and et al., "Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models," *Electronics*, vol. 9, 2020.

[149] S. Trzeciak, R. P. Dellinger, M. E. Chansky, and et al., "Serum lactate as a predictor of mortality in patients with infection," *Intensive Care Medicine*, vol. 33, pp. 970–977, 2007.

[150] D. Misra, V. Avula, D. M. Wolk, and et al., "Early detection of septic shock onset using interpretable machine learners," *Journal of Clinical Medicine*, vol. 10, p. 301, 2021.

[151] C. Kok, V. Jahmunah, S. L. Oh, and et al., "Automated prediction of sepsis using temporal convolutional network," *Computers in Biology and Medicine*, vol. 127, p. 103957, 2020.

[152] Z. M. Ibrahim, H. Wu, A. Hamoud, and et al., "On classifying sepsis heterogeneity in the icu: insight using machine learning," *Journal of the American Medical Informatics Association*, vol. 27, pp. 437–443, 2020.

[153] J. S. Calvert, D. A. Price, U. K. Chettipally, and et al., "A computational approach to early sepsis detection," *Computers in Biology and Medicine*, vol. 74, pp. 69–73, 2016.

[154] Q. Mao, M. Jay, J. L. Hoffman, and et al., "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu," *BMJ Open*, vol. 8, 2018.

[155] K.-C. Yuan, L.-W. Tsai, K.-H. Lee, and et al., "The development of an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit," *International Journal of Medical Informatics*, vol. 141, p. 104176, 2020.

[156] E. Bloch, T. Rotem, J. Cohen, and et al., "Machine learning models for analysis of vital signs dynamics: A case for sepsis onset prediction," *Journal of Healthcare Engineering*, vol. 2019, pp. 1–11, 2019.

[157] M. Saqib, Y. Sha, and M. D. Wang, "Early prediction of sepsis in emr records using traditional ml techniques and deep learning lstm networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 4038–4041.

[158] G. Wardi, M. Carlile, A. Holder, and et al., "Predicting progression to septic shock in the emergency department using an externally generalizable machine learning algorithm," 2020, medRxiv preprint.

[159] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in Biology and Medicine*, vol. 89, pp. 248–255, 2017.

[160] R. Liu, J. L. Greenstein, S. J. Granite, and et al., "Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the icu," *Scientific Reports*, vol. 9, 2019.

[161] A. Johnson, L. Bulgarelli, T. Pollard, and et al., "Mimic-iv," 2021.

[162] M. Singer, C. S. Deutschman, C. W. Seymour, and et al., "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, pp. 801–810, 2016.

[163] T. Kawasaki, "Update on pediatric sepsis: a review," *Journal of Intensive Care*, vol. 5, pp. 1–12, 2017.

[164] S. Inoue, M. Egi, J. Kotani, and et al., "Accuracy of blood-glucose measurements using glucose meters and arterial blood gas analyzers in critically ill adult patients: systematic review," *Critical Care*, vol. 17, p. R48, 2013.

[165] P. Madley-Dowd, R. Hughes, K. Tilling, and et al., "The proportion of missing data should not be used to guide decisions on multiple imputation," *Journal of Clinical Epidemiology*, vol. 110, pp. 63–73, 2019.

[166] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[167] L. M. Stevens, B. J. Mortazavi, R. C. Deo, L. Curtis, and D. P. Kao, "Recommendations for reporting machine learning analyses in clinical research," *Circulation: Cardiovascular Quality and Outcomes*, vol. 13, no. 10, p. e006556, 2020.

[168] S. Shekhar, A. Bansode, and A. Salim, "A comparative study of hyper-parameter optimization tools," in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2021, pp. 1–6.

[169] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, vol. 18, no. 185, pp. 1–52, 2018.

[170] P. Venkatesan, "Treatment response classification in randomized clinical trials: a decision tree approach," *Indian Journal of Science and Technology*, vol. 6, no. 1, pp. 1–6, 2013.

[171] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[172] A. Bhattacharya, *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022.

[173] D. P. Kao, J. D. Lewsey, I. S. Anand, B. M. Massie, M. R. Zile, P. E. Carson, R. S. McKelvie, M. Komajda, J. J. McMurray, and J. Lindenfeld, "Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response," *European journal of heart failure*, vol. 17, no. 9, pp. 925–935, 2015.

[174] M. Faisal, A. Scally, D. Richardson, and et al., "Development and external validation of an automated computer-aided risk score for predicting sepsis in emergency medical admissions using the patient's first electronically recorded vital signs and blood test results," *Critical Care Medicine*, vol. 46, pp. 612–618, 2018.

[175] N. Farzan, S. Vahabi, S. S. Hashemi Madani, and et al., "Evaluating characteristics associated with the mortality among invasive ventilation covid-19 patients," *Annals of Medicine and Surgery*, vol. 69, p. 102832, 2021.

[176] R. Sonneville, F. Verdonk, C. Rauturier, and et al., "Understanding brain dysfunction in sepsis," *Annals of Intensive Care*, vol. 3, 2013.

[177] M. Ebersoldt, T. Sharshar, and D. Annane, "Sepsis-associated delirium," *Intensive Care Medicine*, vol. 33, pp. 941–950, 2007.

[178] J. Allardet-Servent, J.-M. Forel, A. Roch, and et al., "Fio2 and acute respiratory distress syndrome definition during lung protective ventilation," *Critical Care Medicine*, vol. 37, pp. 202–206, 2009.

[179] A. M. Drewry, N. Samra, L. P. Skrupky, and et al., "Persistent lymphopenia after diagnosis of sepsis predicts mortality," *Shock*, vol. 42, pp. 383–391, 2014.

[180] UK Government, "Covid-19 - time from symptom onset until death in uk hospitalised patients," https://www.gov.uk/government/publications/covid-19-time-from-symptom-onset-until-death-in-uk-hospitalised-patients, 2020, accessed: 2023-03-04.

[181] W. H. Organisation. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/coronavirus-disease-(covid-19)

[182] R. Filip, R. Gheorghita Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global challenges to public health care systems during the COVID-19 pandemic: A review of pandemic measures and problems," *J. Pers. Med.*, vol. 12, no. 8, p. 1295, Aug. 2022.

[183] M. Grut, G. de Wildt, J. Clarke, S. Greenfield, and A. Russell, "Primary health care during the COVID-19 pandemic: A qualitative exploration of the challenges and changes in practice experienced by GPs and GP trainees," *PLoS One*, vol. 18, no. 2, p. e0280733, Feb. 2023.

[184] R. C. of General Practitioners, "The future role of remote consultations & patient 'triage'," https://www.rcgp.org.uk/getmedia/72f052b6-3227-48b4-87c9-b05ff90517c4/future-role-of-remote-consultations-patient-triage.pdf, 2021, general practice COVID-19 recovery.

[185] Q. Luo, W. Zeng, M. Chen, G. Peng, X. Yuan, and Q. Yin, "Self-attention and transformers: Driving the evolution of large language models," in *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, 2023, pp. 401–405.

[186] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.

[187] B. Hu, Y. Han, W. Zhang, Q. Zhang, W. Gu, J. Bi, B. Chen, and L. Xiao, "A prediction approach to COVID-19 time series with LSTM integrated attention mechanism and transfer learning," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 323, Dec. 2024.

[188] A. Sehanobish, N. Ravindra, and D. van Dijk, "Gaining insight into sars-cov-2 infection and covid-19 severity using self-supervised edge features and graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 6, 2021, pp. 4864–4873.

[189] B. M. Association, "Delivery of healthcare during the pandemic," https://www.bma.org.uk/media/aganhcxj/bma-covid-review-report-3-september-2024.pdf, 2022, accessed: 2023-05-15.

[190] B. Mahboub, M. T. Bataineh, H. Alshraideh, R. Hamoudi, L. Salameh, and A. Shamayleh, "Prediction of covid-19 hospital length of stay and risk of death using artificial intelligence-based modeling," *Frontiers in Medicine*, vol. 8, p. 389, 5 2021.

[191] G. Wu, P. Yang, Y. Xie, H. C. Woodruff, X. Rao, J. Guiot, A.-N. Frix, R. Louis, M. Moutschen, J. Li, J. Li, C. Yan, D. Du, S. Zhao, Y. Ding, B. Liu, W. Sun, F. Albarello, A. D'Abramo, V. Schininà, E. Nicastri, M. Occhipinti, G. Barisione, E. Barisione, I. Halilaj, P. Lovinfosse, X. Wang, J. Wu, and P. Lambin, "Development of a clinical decision support system for severity risk prediction and triage of covid-19 patients at hospital admission: an international multicenter study," *European Respiratory Journal*, p. 2001104, 7 2020.

[192] D. Assaf, Y. Gutman, Y. Neuman, G. Segal, S. Amit, S. Gefen-Halevi, N. Shilo, A. Epstein, R. Mor-Cohen, A. Biber, G. Rahav, I. Levy, and A. Tirosh, "Utilization of machine-learning models to accurately predict the risk for critical covid-19," *Internal and Emergency Medicine*, vol. 15, pp. 1435–1443, 11 2020.

[193] A. Stachel, K. Daniel, D. Ding, F. Francois, M. Phillips, and J. Lighter, "Development and validation of a machine learning model to predict mortality risk in patients with covid-19," *BMJ Health & Care Informatics*, vol. 28, p. e100235, 5 2021.

[194] F. Rahimian, G. Salimi-Khorshidi, A. H. Payberah, J. Tran, R. A. Solares, F. Raimondi, M. Nazarzadeh, D. Canoy, and K. Rahimi, "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records," *PLOS Medicine*, vol. 15, p. e1002695, 11 2018.

[195] O. Noy, D. Coster, M. Metzger, I. Atar, S. Shenhar-Tsarfaty, S. Berliner, G. Rahav, O. Rogowski, and R. Shamir, "A machine learning model for predicting deterioration of covid-19 inpatients," *Scientific Reports*, vol. 12, p. 2630, 12 2022.

[196] S. Wollenstein-Betech, C. G. Cassandras, and I. C. Paschalidis, "Personalized predictive models for symptomatic covid-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an icu or ventilator." *medRxiv : the preprint server for health sciences*, 5 2020.

[197] Y. Fu, W. Zhong, T. Liu, J. Li, K. Xiao, X. Ma, L. Xie, J. Jiang, H. Zhou, R. Liu, and W. Zhang, "Early prediction model for critical illness of hospitalized covid-19 patients based on machine learning techniques," *Frontiers in Public Health*, vol. 10, 5 2022.

[198] A. A. Willette, S. A. Willette, Q. Wang, C. Pappas, B. S. Klinedinst, S. Le, B. Larsen, A. Pollpeter, T. Li, J. P. Mochel, K. Allenspach, N. Brenner, and T. Waterboer, "Using machine learning to predict covid-19 infection and severity risk among 4510 aged adults: a uk biobank cohort study," *Scientific Reports*, vol. 12, p. 7736, 12 2022.

[199] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," *CoRR*, vol. 56, 11 2015.

[200] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism,"

in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 3512–3520.

[201] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 299–309, 1 2019.

[202] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 65–74.

[203] Z. C. Lipton, D. C. Kale, C. P. Elkan, and R. C. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *CoRR*, vol. abs/1511.03677, 2016.

[204] I. Petersen, I. Douglas, and H. Whitaker, "Self controlled case series methods: an alternative to standard epidemiological study designs," *BMJ*, p. i4515, 9 2016.

[205] W. L. Cava, C. Bauer, J. H. Moore, and S. A. Pendergrass, "Interpretation of machine learning predictions for patient outcomes in electronic health records." *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2019, pp. 572–581, 2019.

[206] E. Alpaydın, "Assessing and comparing classification algorithms," 2001.

[207] G. Tsang and X. Xie, "Deep learning based sepsis intervention: The modelling and prediction of severe sepsis onset," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 8671–8678.

[208] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, and Z. Peng, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in wuhan, china." *JAMA*, vol. 323, pp. 1061–1069, 2020.

[209] A. Benksim, R. Ait Addi, and M. Cherkaoui, "Vulnerability and fragility expose older adults to the potential dangers of COVID-19 pandemic," *Iran. J. Public Health*, vol. 49, no. Suppl 1, pp. 122–124, Oct. 2020.

[210] A. Sharma and S. Sharma, "Impact of COVID-19 pandemic on the elderly in the united kingdom: A review study," *J. Family Med. Prim. Care*, vol. 13, no. 8, pp. 2826–2833, Aug. 2024.

[211] C. Munteanu and B. Schwartz, "The relationship between nutrition and the immune system," *Front. Nutr.*, vol. 9, p. 1082500, Dec. 2022.

[212] M. F. Dinevari, M. H. Somi, E. S. Majd, M. A. Farhangi, and Z. Nikniaz, "Anemia predicts poor outcomes of covid-19 in hospitalized patients: a prospective study in iran," *BMC Infectious Diseases*, vol. 21, p. 170, 12 2021.

[213] S. Agarwal and K. Karkouti, "The relationship between anaemia and poor outcomes: let's get to the meat of the matter," *Anaesthesia*, vol. 76, no. 10, pp. 1300–1303, Oct. 2021.

[214] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari, "Clinical text summarization: Adapting large language models can outperform human experts," *Res. Sq.*, Oct. 2023.

[215] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang, J. Gao, Y.-G.-S. Wang, J.-M. Ji, Z. Qiu, M. Li, C. Qian, T. Guo, S. Ma, Z. Wang, Z. Guo, Y. Lei, C. Shao, W. Wang, H. Fan, and Y.-D. Tang, "The application of large language models in medicine: A scoping review," *iScience*, vol. 27, no. 5, p. 109713, May 2024.

[216] M. Rybinski, W. Kusa, S. Karimi, and A. Hanbury, "Learning to match patients to clinical trials using large language models," *J. Biomed. Inform.*, vol. 159, no. 104734, p. 104734, Nov. 2024.

[217] S. Aydin, M. Karabacak, V. Vlachos, and K. Margetis, "Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support," *Front. Med. (Lausanne)*, vol. 12, p. 1527864, Jan. 2025.

[218] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang, J. Gao, Y.-G.-S. Wang, J.-M. Ji, Z. Qiu, M. Li, C. Qian, T. Guo, S. Ma, Z. Wang, Z. Guo, Y. Lei, C. Shao, W. Wang, H. Fan, and Y.-D. Tang, "The application

of large language models in medicine: A scoping review," *iScience*, vol. 27, no. 5, p. 109713, May 2024.

[219] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, S. J. Wagner, and J. N. Kather, "The future landscape of large language models in medicine," *Commun. Med. (Lond.)*, vol. 3, no. 1, p. 141, Oct. 2023.

[220] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[221] J. G. Diaz Ochoa and F. E. Mustafa, "Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses," *Artif. Intell. Med.*, vol. 131, no. 102359, p. 102359, Sep. 2022.

[222] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[223] Z. Liu, X. Li, H. Peng, L. He, and P. S. Yu, "Heterogeneous similarity graph neural network on electronic health records," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2020.

[224] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI," *Inf. Fusion*, vol. 71, pp. 28–37, Jul. 2021.

[225] D. He, R. Wang, Z. Xu, J. Wang, P. Song, H. Wang, and J. Su, "The use of artificial intelligence in the treatment of rare diseases: A scoping review," *Intractable Rare Dis. Res.*, vol. 13, no. 1, pp. 12–22, Feb. 2024.

[226] T. Groza, C.-H. Chan, D. A. Pearce, and G. Baynam, "Realising the potential impact of artificial intelligence for rare diseases – a framework," *Rare*, vol. 3, no. 100057, p. 100057, 2025.

[227] R. Filip, R. Gheorghita Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global challenges to public health care systems during the covid-19 pandemic: A

review of pandemic measures and problems," *Journal of Personalized Medicine*, vol. 12, no. 8, p. 1295, Aug 2022. [Online]. Available: http://dx.doi.org/10.3390/jpm12081295

[228] A. Fernández, S. Horng, Q. Wells, D. Vawdrey, X. Li, S. E. Perlman, and B. S. Glicksberg, "Prediction models to estimate hospitalization, intensive care unit admission, and mortality risk using machine learning and natural language processing of electronic health records in covid-19 patients: protocol for a retrospective cohort study," *JMIR Research Protocols*, vol. 10, no. 3, p. e26084, 2021.

[229] J. Shang, Y. Sun, N. Zhang, X. Chen, and N. Bao, "Machine learning methods for predicting outcomes of covid-19 patients," *PLoS ONE*, vol. 15, no. 11, p. e0242303, 2020.

[230] K. Gong, D. Wu, C. D. Arru, F. Homayounieh, N. Neumark, J. Guan, V. Buch, K. Kim, B. C. Bizzo, H. Ren, W. Y. Tak, S. Y. Park, Y. R. Lee, M. K. Kang, J. G. Park, A. Carriero, L. Saba, M. Masjedi, H. Talari, R. Babaei, H. K. Mobin, S. Ebrahimian, N. Guo, S. R. Digumarthy, I. Dayan, M. K. Kalra, and Q. Li, "A multi-center study of covid-19 patient prognosis using deep learning-based ct image analysis and electronic health records," *European Journal of Radiology*, vol. 139, p. 109583, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0720048X21000632

[231] J. Watson, B. D. Nicholson, W. Hamilton, and S. Price, "Identifying clinical features in primary care electronic health record studies: methods for codelist development," *BMJ Open*, vol. 7, no. 11, p. e019637, Nov. 2017. [Online]. Available: https://doi.org/10.1136/bmjopen-2017-019637

[232] J. S. Tulloch, M. B. Beadsworth, R. Vivancos, A. D. Radford, J. C. Warner, and R. M. Christley, "GP coding behaviour for non-specific clinical presentations: a pilot study," *BJGP Open*, vol. 4, no. 3, p. bjgpopen20X101050, Jul. 2020. [Online]. Available: https://doi.org/10.3399/bjgpopen20x101050

[233] K. Hur, J. Lee, J. Oh, W. Price, Y. Kim, and E. Choi, "Unifying heterogeneous electronic health records systems via text-based code embedding," in *Proceedings of the Conference on Health, Inference, and Learning*, ser. Proceedings of Machine Learning Research, G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, Eds., vol. 174. PMLR, 07–08 Apr 2022, pp. 183–203. [Online]. Available: https://proceedings.mlr.press/v174/hur22a.html

[234] J. G. Diaz Ochoa and F. E. Mustafa, "Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses," *Artificial Intelligence in Medicine*, vol. 131, p. 102359, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365722001245

[235] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. Jim Zheng, and K. Roberts, "Deep representation learning of patient data from electronic health records (ehr): A systematic review," *Journal of Biomedical Informatics*, vol. 115, p. 103671, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420302999

[236] S. Ramchand, G. Tsang, D. Cole, and X. Xie, "Retainext: Enhancing rare event detection and improving interpretability of health records using temporal neural networks," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2022, pp. 01–06.

[237] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[238] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: Graph-based attention model for healthcare representation learning," 2017.

[239] E. Choi, Z. Xu, Y. Li, M. W. Dusenberry, G. Flores, Y. Xue, and A. M. Dai, "Learning the graphical structure of electronic health records with graph convolutional transformer," 2020.

[240] H. Lu and S. Uddin, "A weighted patient network-based framework for predicting chronic diseases using graph neural networks," *Scientific Reports*, vol. 11, no. 1, p. 22607, 2021.

[241] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[242] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," 2020.

[243] J. Smith and E. Johnson, "Predictive modeling in medicine: Current trends and limitations," *Journal of Medical Informatics*, vol. 45, no. 3, pp. 234–245, 2023.

[244] D. Brown and S. Lee, "The gap in risk assessment models: A comprehensive review," *Health Informatics Review*, vol. 18, no. 2, pp. 112–128, 2023.

[245] M. Garcia and R. Wilson, "Advancements and challenges in medical predictive modeling," in *Proceedings of the International Conference on Medical AI*. International Medical AI Society, 2023, pp. 78–85.

[246] E. Taylor and M. Anderson, "The role of taxonomies in medical graph constructions: A critical analysis," *Journal of Biomedical Data Science*, vol. 12, no. 4, pp. 345–360, 2023.

[247] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "Behrt: Transformer for electronic health records," *Scientific Reports*, vol. 10, no. 1, p. 7155, 2020.

[248] R. A. Lyons, K. H. Jones, G. John, C. J. Brooks, J.-P. Verplancke, D. V. Ford, G. Brown, and K. Leake, "The sail databank: linking multiple health and social care datasets," *BMC Medical Informatics and Decision Making*, vol. 9, p. 3, 12 2009.

[249] Y. Jung and J. Hu, "AK-fold averaging cross-validation procedure," *Journal of Nonparametric Statistics*, vol. 27, no. 2, pp. 167–179, feb 2015. [Online]. Available: https://doi.org/10.1080/10485252.2015.1010532

[250] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962v2*, 2019.

[251] P. Dubach, J. Myers, P. Bonetti, T. Schertler, V. Froelicher, D. Wagner, M. Scheidegger, M. Stuber, R. Luchinger, J. Schwitter *et al.*, "Effects of bisoprolol fumarate on left ventricular size, function, and exercise capacity in patients with heart failure: analysis with magnetic resonance myocardial tagging," *American Heart Journal*, vol. 143, no. 4, pp. 676–683, 2002.

[252] J. A. Dickinson, "Lesser-spotted zebras: Their care and feeding," *Can. Fam. Physician*, vol. 62, no. 8, pp. 620–621, Aug. 2016.

[253] D. P. Germain, "Fabry disease," *Orphanet journal of rare diseases*, vol. 5, no. 1, pp. 1–49, 2010.

[254] R. Garg, S. Dong, S. Shah, and S. R. Jonnalagadda, "A bootstrap machine learning approach to identify rare disease patients from electronic health records," *arXiv preprint arXiv:1609.01586*, 2016.

[255] B. Hoffmann and E. Mayatepek, "Fabry disease—often seen, rarely diagnosed," *Deutsches Ärzteblatt international*, vol. 106, no. 26, p. 440, 2009.

[256] U. Ramaswami, C. Whybra, R. Parini, G. Pintos-Morell, A. Mehta, G. Sunder-Plassmann, U. Widmer, M. Beck, and F. E. Investigators, "Clinical manifestations of fabry disease in children: data from the fabry outcome survey," *Acta paediatrica*, vol. 95, no. 1, pp. 86–92, 2006.

[257] A. Mehta, R. Ricci, U. Widmer, F. Dehout, A. Garcia de Lorenzo, C. Kampmann, A. Linhart, G. Sunder-Plassmann, M. Ries, and M. Beck, "Fabry disease defined: baseline clinical manifestations of 366 patients in the fabry outcome survey," *European journal of clinical investigation*, vol. 34, no. 3, pp. 236–242, 2004.

[258] D. A. Hughes, P. Aguiar, O. Lidove, K. Nicholls, A. Nowak, M. Thomas, R. Torra, B. Vujkovac, M. L. West, and S. Feriozzi, "Do clinical guidelines facilitate or impede drivers of treatment in fabry disease?" *Orphanet Journal of Rare Diseases*, vol. 17, no. 1, pp. 1–15, 2022.

[259] H. Uryu, O. Migita, M. Ozawa, C. Kamijo, S. Aoto, K. Okamura, F. Hasegawa, T. Okuyama, M. Kosuga, and K. Hata, "Automated urinary sediment detection for fabry disease using deep-learning algorithms," *Molecular Genetics and Metabolism Reports*, vol. 33, p. 100921, 2022.

[260] J. L. Jefferies, A. K. Spencer, H. A. Lau, M. W. Nelson, J. D. Giuliano, J. W. Zabinski, C. Boussios, G. Curhan, R. E. Gliklich, and D. G. Warnock, "A new approach to identifying patients with elevated risk for fabry disease using a machine learning algorithm," *Orphanet Journal of Rare Diseases*, vol. 16, pp. 1–8, 2021.

[261] A. A. Michalski, K. Lis, J. Stankiewicz, S. M. Kloska, A. Sycz, M. Dudziński, K. Muras-Szwedziak, M. Nowicki, S. Bazan-Socha, M. J. Dabrowski *et al.*, "Supporting the diag-

nosis of fabry disease using a natural language processing-based approach," *Journal of Clinical Medicine*, vol. 12, no. 10, p. 3599, 2023.

[262] A. R. S. Moura, "Detecting important electrocardiogram characteristics for the diagnosis of fabry disease via statistics and machine learning techniques," Ph.D. dissertation, universidade do minho, 2022.

[263] J. Yim, O. Yau, D. F. Yeung, and T. S. Tsang, "Fabry cardiomyopathy: current practice and future directions," *Cells*, vol. 10, no. 6, p. 1532, 2021.

[264] A. Linhart, D. P. Germain, I. Olivotto, M. M. Akhtar, A. Anastasakis, D. Hughes, M. Namdar, M. Pieroni, A. Hagege, F. Cecchi *et al.*, "An expert consensus document on the management of cardiovascular manifestations of fabry disease," *European journal of heart failure*, vol. 22, no. 7, pp. 1076–1096, 2020.

[265] R. Clark, "Perspectives on machine learning and artificial intelligence from trainee radiologists," Swansea University Swansea, 2021.

[266] M. Teng, R. Singla, O. Yau, D. Lamoureux, A. Gupta, Z. Hu, R. Hu, A. Aissiou, S. Eaton, C. Hamm *et al.*, "Health care students' perspectives on artificial intelligence: countrywide survey in canada," *JMIR medical education*, vol. 8, no. 1, p. e33390, 2022.

[267] P. Elliott, R. Baker, F. Pasquale, G. Quarta, H. Ebrahim, A. B. Mehta, and D. A. Hughes, "Prevalence of anderson–fabry disease in patients with hypertrophic cardiomyopathy: the european anderson–fabry disease survey," *Heart*, vol. 97, no. 23, pp. 1957–1960, 2011.

[268] M. S. Maron, W. Xin, K. B. Sims, R. Butler, T. S. Haas, E. J. Rowin, R. J. Desnick, and B. J. Maron, "Identification of fabry disease in a tertiary referral cohort of patients with hypertrophic cardiomyopathy," *The American Journal of Medicine*, vol. 131, no. 2, pp. 200–e1, 2018.

[269] L. Monserrat, J. R. Gimeno-Blanes, F. Marín, M. Hermida-Prieto, A. García-Honrubia, I. Pérez, X. Fernández, R. de Nicolas, G. de la Morena, E. Payá *et al.*, "Prevalence of fabry disease in a cohort of 508 unrelated patients with hypertrophic cardiomyopathy," *Journal of the American College of Cardiology*, vol. 50, no. 25, pp. 2399–2403, 2007.

[270] S. Sen-Chowdhry, D. Jacoby, J. C. Moon, and W. J. McKenna, "Update on hypertrophic cardiomyopathy and a guide to the guidelines," *Nature Reviews Cardiology*, vol. 13, no. 11, pp. 651–675, 2016.

[271] T. H. R. Costa, J. A. de Figueiredo Neto, A. E. F. de Oliveira, M. d. F. L. e Maia, and A. L. de Almeida, "Association between chronic apical periodontitis and coronary artery disease," *Journal of endodontics*, vol. 40, no. 2, pp. 164–167, 2014.

[272] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of hyperparameters of machine learning algorithms," *Journal of Machine Learning Research*, vol. 20, no. 53, pp. 1–32, 2019.

[273] A. Mehta and U. Widmer, "Natural history of fabry disease," in *Fabry Disease: Perspectives from 5 Years of FOS*. Oxford: Oxford PharmaGenesis, 2006.

[274] A. A. Michalski, K. Lis, J. Stankiewicz, S. M. Kloska, A. Sycz, M. Dudziński, K. Muras-Szwedziak, M. Nowicki, S. Bazan-Socha, M. J. Dabrowski, and G. W. Basak, "Supporting the diagnosis of fabry disease using a natural language processing-based approach," *J. Clin. Med.*, vol. 12, no. 10, May 2023.

[275] J. Jefferies, P. Aguiar, G. Biondetti, D. Warnock, S. Kallish, M. Nelson, J. Giuliano, J. Zabinksi, C. Boussios, G. Curhan *et al.*, "Estimation of arrhythmia risk in patients with fabry disease using a machine learning model," *The Journal of Heart and Lung Transplantation*, vol. 42, no. 4, p. S331, 2023.

[276] J. Jefferies, S. Kallish, G. Biondetti, P. Aguiar, M. Nelson, J. Giuliano, J. Zabinksi, C. Boussios, G. Curhan, J. Bandaria *et al.*, "Estimation of stroke risk in patients with fabry disease using a machine learning model," *The Journal of Heart and Lung Transplantation*, vol. 42, no. 4, p. S332, 2023.

[277] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–7.

[278] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu *et al.*, "Prompt engineering for healthcare: Methodologies and applications," *arXiv preprint arXiv:2304.14670*, 2023.

[279] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[280] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[281] S. Guthals and P. Haack, *GitHub for dummies*. John Wiley & Sons, 2019.

[282] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," *ArXiv*, vol. abs/2310.06825, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263830494

[283] A. Gramegna and P. Giudici, "Shap and lime: an evaluation of discriminative power in credit risk," *Frontiers in Artificial Intelligence*, vol. 4, p. 752558, 2021.

Appendix A

Supplementary Materials

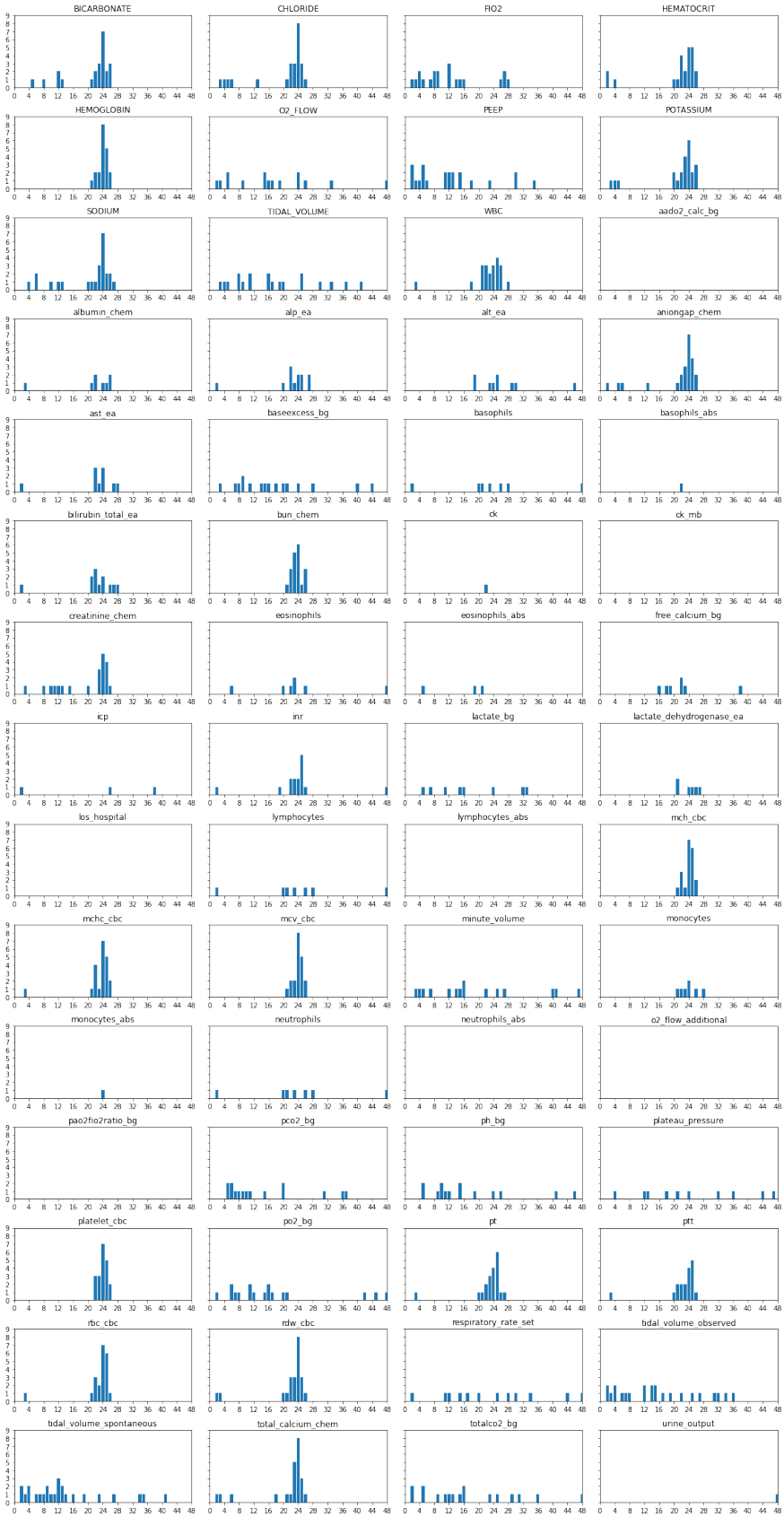


Figure A.1: Frequency of measurements for each feature in Sepsis patients.

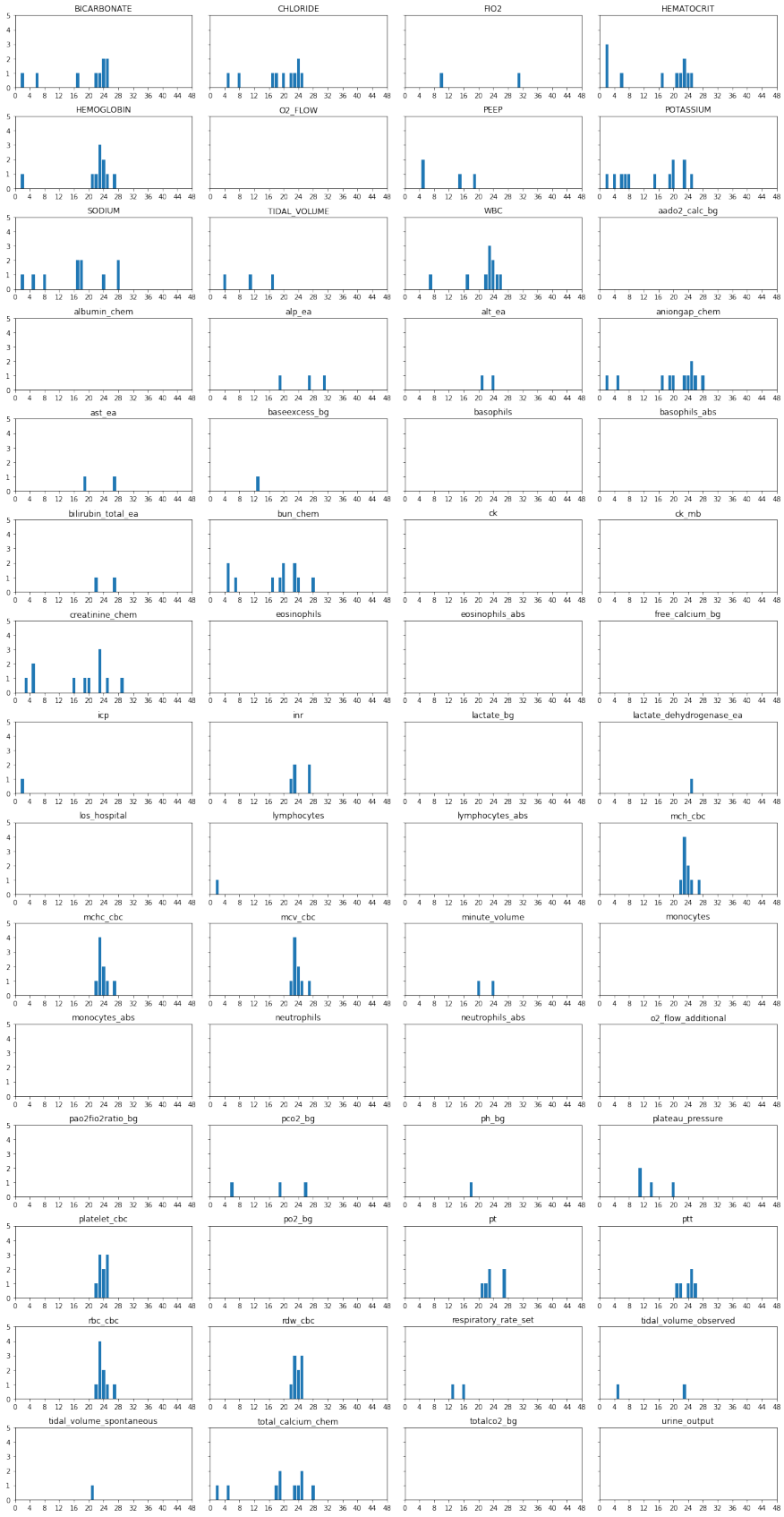


Figure A.2: Frequency of measurements for each feature in control patients.

Clinical Marker	Category	Variable Type	Minimum	Maximum	Units	Missingness	Step 2 of Imputation strategy	Step 3 of Imputation Strategy - Static	Step 3 of Imputation Strategy - Dynamic
Height	Administrative	Numerical	50	275	cm	MCAR	Forward Fill	Median	Not Applicable
Weight	Administrative	Numerical	10	650	kg	MCAR	Forward Fill	Median	Not Applicable
Age	Administrative	Numerical	18	100	Years	No Missingness	Forward Fill	Median	Not Applicable
Ethnicity	Administrative	Categorical	White; Asian; Black/African American; Hispanic/Latino; Other			No Missingness	-	Label as "Unknown"	
Gender	Administrative	Categorical	Male/Female			No Missingness	-	Label as "Unknown"	
Length of Hospital Stay	Administrative	Numerical	0	-	Hours	No Missingness	-	-	-
Invasive Ventilation	Ventilator type	Binary	0	1	-	-	-	-	-
GCS - Verbal Response	Clinical Observations	Numerical	0	5	-	MCAR	Interpolate	Median	Forward Fill
Tracheal ventilation	Ventilator type	Binary	0	1	-	-	-	-	-
Unable to take GCS	Clinical Observations	Numerical	0	1	-	MCAR	Leave Blank		
Fraction of inspired oxygen	Laboratory/Ventilator	Numerical	21	100	%	MNAR	Interpolate	Median	Forward Fill
Respiratory Rate (Spontaneous)	Clinical settings	Numerical	0	30	breaths/min	MCAR	Interpolate	Median	Forward Fill
Lymphocytes	Blood Differentials	Numerical	0	100	%	MNAR	Interpolate	Median	Forward Fill
PaO2	Blood Gases	Numerical	0	600	mm Hg	MNAR	Interpolate	Median	Forward Fill
Lactate	Blood Gases	Numerical	0.05	37	mmol/L	MCAR	Interpolate	Median	Forward Fill
Tidal Volume (Observed)	Clinical settings	Numerical	100	6500	mL	MNAR	Interpolate	Median	Forward Fill
Severe Liver Disease	Comorbidity	Binary	0	1	-	-	-	-	-
Tidal Volume (Spontaneous)	Ventilator setting	Numerical	100	6500	mL	MNAR	Interpolate	Median	Forward Fill
Intracranial Pressure	Clinical Observations	Numerical	1	50	mmHg	MNAR	Interpolate	Median	Forward Fill
Respiratory Rate (Set)	Clinical settings	Numerical	0	30	breaths/min	MNAR	Interpolate	Median	Forward Fill

Table A.1: Clinical markers with data type and category, clinical ranges and measurement units, type of missingness and steps of imputation used.

Sepsis Diagnosis	Hours Earlier	Scaled & Encoded (1)	Interpolated (2)	Forward Fill or Median (3)	Full Set of Features					N. Feat Selected	Importance Threshold	Selected Set of Features				
					Accuracy	F1 Score	AUROC	Sensitivity	Specificity			Accuracy	F1 Score	AUROC	Sensitivity	Specificity
Onset	0	✓	✗	✗	91.98%	89.13%	90.79%	85.67%	95.91%	13	1.2%	75.61%	63.36%	71.71%	54.96%	88.47%
		✓	✓	✗	96.35%	95.05%	95.47%	91.74%	99.21%	13	1.3%	82.77%	73.86%	79.13%	63.73%	94.53%
		✓	✓	✓	99.58%	99.45%	99.53%	99.33%	99.74%	14	1.3%	92.71%	89.87%	91.17%	84.66%	97.68%
	6	✓	✗	✗	92.79%	90.62%	91.88%	87.42%	96.35%	16	1.2%	75.55%	64.75%	72.31%	56.38%	88.23%
		✓	✓	✗	96.07%	94.89%	95.31%	91.55%	99.07%	16	1.2%	83.40%	75.84%	80.36%	65.40%	95.32%
		✓	✓	✓	99.07%	98.83%	98.93%	98.25%	99.62%	13	1.4%	91.86%	89.10%	90.46%	83.55%	97.36%
	12	✓	✗	✗	94.88%	93.57%	94.31%	91.24%	97.39%	16	1.2%	75.23%	65.19%	72.37%	56.80%	87.95%
		✓	✓	✗	96.02%	94.95%	95.34%	91.61%	99.06%	16	1.2%	82.95%	75.80%	80.23%	65.36%	95.10%
		✓	✓	✓	99.05%	98.82%	98.92%	98.23%	99.61%	14	1.4%	91.39%	88.78%	90.15%	83.36%	96.94%
Sepsis 3	0	✓	✗	✗	89.91%	87.34%	89.05%	84.10%	94.01%	16	1.2%	75.02%	66.68%	72.87%	60.41%	85.33%
		✓	✓	✗	95.80%	94.75%	95.18%	91.61%	98.76%	16	1.2%	84.00%	78.32%	81.92%	69.91%	93.93%
		✓	✓	✓	99.40%	99.27%	99.35%	99.07%	99.62%	12	1.3%	93.14%	91.29%	92.22%	86.88%	97.56%
	6	✓	✗	✗	93.29%	92.04%	92.89%	89.94%	95.84%	13	1.2%	73.06%	65.82%	71.50%	60.11%	82.90%
		✓	✓	✗	95.57%	94.68%	95.08%	91.50%	98.65%	15	1.2%	82.47%	76.94%	80.70%	67.79%	93.60%
		✓	✓	✓	98.88%	98.70%	98.78%	98.04%	99.53%	16	1.3%	93.08%	91.57%	92.36%	87.04%	97.67%
	12	✓	✗	✗	93.69%	92.69%	93.37%	90.59%	96.14%	13	1.2%	72.60%	66.08%	71.33%	60.45%	82.20%
		✓	✓	✗	95.46%	94.70%	95.06%	91.50%	98.62%	16	1.2%	81.55%	76.34%	80.08%	67.26%	92.89%
		✓	✓	✓	98.86%	98.70%	98.77%	98.01%	99.53%	14	1.3%	91.90%	90.26%	91.19%	85.21%	97.17%
Suspected Infection	0	✓	✗	✗	96.65%	93.93%	95.49%	92.87%	98.11%	16	1.2%	83.11%	65.73%	75.43%	58.02%	92.83%
		✓	✓	✗	98.62%	97.49%	97.79%	95.90%	99.68%	15	1.3%	91.93%	83.95%	86.91%	75.53%	98.29%
		✓	✓	✓	99.70%	99.46%	99.57%	99.29%	99.86%	15	1.0%	97.51%	95.39%	96.03%	92.71%	99.35%
	6	✓	✗	✗	95.97%	92.84%	94.48%	90.98%	97.98%	16	1.1%	82.95%	66.39%	75.69%	58.62%	92.75%
		✓	✓	✗	98.55%	97.43%	97.69%	95.68%	99.71%	14	1.3%	90.44%	81.34%	85.02%	72.20%	97.84%
		✓	✓	✓	99.61%	99.33%	99.42%	98.95%	99.88%	16	1.1%	97.20%	94.99%	95.62%	91.87%	99.37%
	12	✓	✗	✗	97.00%	94.85%	96.07%	93.79%	98.34%	16	1.2%	82.55%	66.46%	75.61%	58.75%	92.47%
		✓	✓	✗	98.46%	97.34%	97.58%	95.44%	99.73%	16	1.2%	91.45%	83.86%	86.76%	75.33%	98.19%
		✓	✓	✓	99.56%	99.25%	99.33%	98.77%	99.89%	14	1.1%	97.04%	94.80%	95.42%	91.48%	99.37%

Table A.2: All Model Performances across diagnosis points, point 0 (sepsis-3 diagnosis time), 6 (6 hours prior to point 0) and 12 (12 hours prior to point 0) and for full and reduced feature sets. The best performances are highlighted.