



IMPROVING THE ROBUSTNESS OF MULTIMODAL AI WITH ASYNCHRONOUS AND MISSING INPUTS

RICARDO BRUNO BARBEIRO DOS SANTOS

Master in Biomedical Engineering

DOCTORATE IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
September, 2024

IMPROVING THE ROBUSTNESS OF MULTIMODAL AI WITH ASYNCHRONOUS AND MISSING INPUTS

RICARDO BRUNO BARBEIRO DOS SANTOS

Master in Biomedical Engineering

Advisers: Hugo Filipe Silveira Gamboa
Full Professor, NOVA School of Science and Technology
André Valério Raposo Carreiro
Senior Scientist, Fraunhofer Portugal AICOS

Examination Committee

Chair: Maria de Fátima Guerreiro da Silva Campos Raposo
Full Professor, NOVA School of Science and Technology

Rapporteurs: Jordi Solé-Casals
Full Professor, Universitat de Vic - Universitat Central de Catalunya

Ana Catarina Fidalgo Barata
Assistant Professor, Instituto Superior Técnico - Universidade de Lisboa

Adviser: André Valério Raposo Carreiro
Senior Scientist, Fraunhofer Portugal AICOS

Member: Pedro Manuel Cardoso Vieira
Associate Professor, NOVA School of Science and Technology

Improving the Robustness of Multimodal AI with Asynchronous and Missing Inputs

Copyright © Ricardo Bruno Barbeiro dos Santos, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

ACKNOWLEDGEMENTS

As I reach the end of this journey, I feel a sense of personal satisfaction and an opportunity to reflect on the intense years that have led me here. It is time to share this success with all the incredible people and institutions who were by my side. Without them, I would not be writing this text today. The journey was challenging and tested me in ways I had not anticipated, but the support of the following allowed me to cross the finish line.

First and foremost, I want to express my gratitude to my two advisers. Professor Hugo Gamboa has been the voice of experience, guiding me on both scientific and personal excellence. His vast knowledge, honesty, and serious approach have been an inspiring lesson in humility. I also owe a huge thank you to Dr. André Carreiro, who has been incredibly important in shaping the researcher I am today. Thank you for your moral support, endless creativity, intelligence, positive outlook, and, foremost, for believing in my potential. I am proud to have been your first PhD student and look forward to seeing the success of those who will follow.

I also thank Fraunhofer Portugal AICOS, my research institution, for giving me this opportunity and providing the conditions necessary to succeed. A particular word of appreciation goes to Liliana Ferreira and Inês Sousa, who demonstrated exemplary leadership and were always concerned about my well-being, supporting my ideas along the way.

During this journey, I was fortunate to embark on two research stays abroad, which enriched me both personally and scientifically. Professor Holger Fröhlich, thank you for welcoming me to Fraunhofer SCAI, involving me in research discussions, and guiding me with fresh ideas and perspectives. Professor Xi Peng, thank you for making my time at the University of Delaware's DeepREAL Lab an unforgettable experience. Thanks to your support, I gained new perspectives on how to organize and approach research.

Several other institutions also played an important role in my success, enabling me to conduct research and grow as a scientist. Thank you to the VOH.CoLAB and Hospital de Santa Marta teams for the close partnership. I am also grateful to the Luso-American Development Foundation for the scholarship that supported my research stay in the USA.

Looking back on the past four years, I am deeply thankful for the remarkable people

who have crossed my path. To the wonderful friends I made in Bonn and Delaware, Jonas, Yasamin, Mario, and all the others, thank you for welcoming me with open arms and making me feel at home.

To my colleagues and friends at work, Bruno, Tomás, Pedro, Marília, Isabel, and everyone else, thank you for making day-to-day life more enjoyable and helping to make the hard times easier to bear. I am confident that our professional paths will cross again, and I look forward to seeing all you achieve.

I cannot forget my dearest friends who have supported me throughout this journey. You have been my rock, bringing me joy and cheering me on. I won't name everyone individually since the last four years have impacted my memory, and I don't want the risk of forgetting anyone, but you know who you are. Thank you for everything, the past shared moments, and all that is yet to come.

Above all, I owe my deepest gratitude to my family. Your love, guidance, and support have been the foundation of my success. Thank you for giving me the tools to pursue my professional and personal goals. Muito muito obrigado.

I cannot finish these acknowledgments without thinking of those who supported me at the start of this journey but are no longer here to share this accomplishment. To Leonardo and Avó Joaquina, thank you for always being there. You will forever have a place in my heart, and I will continue striving to be someone who makes you proud.

ABSTRACT

Integrating Artificial Intelligence (AI) in clinical Decision Support Systems (DSS) can significantly transform healthcare in many ways to improve patient outcomes. Despite these promises, AI-based clinical DSS face critical challenges, particularly in handling complex, asynchronous, and incomplete medical data. This lack of robustness in AI models and issues around insufficient validation and clinician trust pose substantial barriers to effectively deploying these systems in real-world settings.

This thesis addresses these challenges by proposing a holistic framework to develop robust AI-based clinical DSS. The framework is structured around four core strategies: human-centric, data-centric, model-centric, and deployment-centric approaches. Through this framework, we aim to overcome the limitations of current systems and create more valuable and generalizable AI solutions in healthcare.

To demonstrate the value of this framework, we explored three medical use cases, each presenting distinct challenges and technical objectives. First, we developed risk models for predicting postoperative complications using a clinician-in-the-loop approach in cardiothoracic surgery. This work focused on addressing real-world data heterogeneity and missing inputs to improve the quality and relevance of the predictions throughout the patient journey. Second, we explored multimodal data integration within cardiovascular disease classification from electrocardiogram signals. We investigated how different data representations in a multimodal framework can be leveraged to enhance model robustness and identified that multimodal schemes are only sometimes superior. Finally, we researched the challenge of handling numerous asynchronous medical variables in continuous monitoring within critical care. We developed ICU-BERT, a transformer-based model designed to effectively process complex, time-sensitive data, improving real-time decision support. To show its generalization potential, we tested ICU-BERT in various databases and real-world tasks.

This thesis's contributions extend beyond technical advancements; they include strategies to ensure that AI-based clinical DSS can be integrated effectively into clinical practice. By addressing key challenges in AI model robustness, this work lays the foundation for developing trustworthy clinical DSS that can enhance patient care.

Keywords: Clinical Decision Support Systems, Robust Artificial Intelligence, Cardiothoracic Surgery, Electrocardiogram, Intensive Care Unit

RESUMO

A integração de Inteligência Artificial (IA) em Sistemas de Suporte à Decisão (SSD) clínicos pode transformar significativamente a prestação de cuidados de saúde, melhorando os resultados clínicos de diversas formas. Apesar deste potencial, os SSD baseados em IA enfrentam desafios críticos, especialmente no tratamento de dados médicos, que são complexos, assíncronos e incompletos. A falta de robustez nos modelos de IA, associada à insuficiente validação e falta de confiança dos profissionais de saúde, impedem a implementação eficaz destes sistemas em cenários reais.

O presente trabalho aborda esses desafios, propondo uma estratégia holística para o desenvolvimento de SSD clínicos mais robustos baseados em IA. A estratégia apresentada estrutura abordagens técnicas em torno de quatro eixos: metodologias centradas no humano, nos dados, nos modelos e na sua implementação. Com esta estratégia, pretende-se ultrapassar as limitações dos sistemas atuais e criar soluções de IA mais úteis e generalizáveis para os cuidados de saúde.

Para demonstrar a eficácia desta estratégia, explorámos três casos de uso clínicos, cada um apresentando objetivos e desafios técnicos distintos. Em primeiro lugar, desenvolvemos modelos de previsão de risco de complicações pós-operatórias em cirurgia cardiotorácica, envolvendo a participação ativa de profissionais clínicos. Este trabalho focou-se em abordar a recorrente escassez e heterogeneidade de dados em contextos reais, visando melhorar a qualidade e relevância das previsões ao longo do percurso do paciente. De seguida, investigámos a integração de dados multimodais na classificação de doenças cardiovasculares a partir de sinais de eletrocardiograma. Analisámos como diferentes representações de dados num modelo multimodal podem ser aproveitadas para aumentar a robustez, identificando porém que estes métodos nem sempre são superiores. Finalmente, abordámos o desafio de lidar com grandes quantidades de variáveis médicas assíncronas na monitorização contínua em cuidados intensivos. Desenvolvemos o ICU-BERT, um modelo baseado em *transformers* para processar eficazmente dados complexos e temporais, melhorando o suporte à decisão em tempo real. Para demonstrar o seu potencial de generalização, testámos o ICU-BERT em várias bases de dados e tarefas do mundo real.

As contribuições desta tese vão além de avanços técnicos; incluem estratégias para

garantir que SSD clínicos baseados em IA possam ser integrados eficazmente na prática clínica. Ao abordar os principais desafios na robustez dos modelos de IA, este trabalho estabelece as bases para o futuro desenvolvimento de SSD clínicos confiáveis que efetivamente melhorem a prestação de cuidados.

Palavras-chave: Sistemas de Suporte à Decisão Clínicos, Inteligência Artificial Robusta, Cirurgia Cardiotorácica, Eletrocardiograma, Unidade de Cuidados Intensivos

CONTENTS

List of Figures	xii
List of Tables	xiv
Acronyms	xvi
Prologue	1
1 Introduction	3
1.1 Motivation	3
1.2 Research Questions	4
1.3 Objectives and Contributions	5
1.4 Thesis Structure	8
2 Theoretical Background	10
2.1 Machine Learning	10
2.1.1 Classification Models	12
2.2 Deep Learning	13
2.2.1 Deep Learning Strategies	14
2.2.2 Attention Mechanism	16
2.3 Medical Data Processing	17
2.3.1 Missing Data	18
2.3.2 Correlation	20
2.4 The Machine Learning Pipeline in Healthcare	21
2.4.1 Data Collection and Preparation	21
2.4.2 Feature Engineering and Selection	22
2.4.3 Model Development	23
2.4.4 Evaluation	24
2.4.5 Deployment	26
2.5 Final Remarks	26

3	Foundations of Robust Artificial Intelligence in Healthcare	28
3.1	Decision Support Systems in Healthcare	29
3.1.1	Current State	30
3.1.2	Limitations	31
3.1.3	Future Potential	33
3.2	Real World Medical Data	34
3.2.1	Challenges	36
3.3	A Holistic Approach to Robust AI	38
3.4	Ethical and Legal Frameworks for Responsible and Trustworthy AI . . .	41
3.4.1	Ethical Principles and Key Requirements for Trustworthy AI . . .	42
3.4.2	Regulatory Frameworks	44
3.5	Final Remarks	46
4	Robustness in Cardiothoracic Surgery Risk Estimation: Addressing Data Heterogeneity and Missing Inputs	48
4.1	Introduction	48
4.2	Case Study: CardioFollow.AI	49
4.2.1	Remote Patient Monitoring	50
4.2.2	Data Collection	50
4.2.3	Artificial Intelligence Tasks	53
4.3	Developing Risk Models for Cardiothoracic Surgery	54
4.3.1	Objectives and Significance	54
4.3.2	Literature Review	55
4.3.3	Methodology	57
4.3.4	Experimental Results	61
4.3.5	Discussion	65
4.3.6	Real-World Validation	68
4.3.7	Limitations	69
4.4	Handling Missing Data from the Real World	70
4.4.1	Objectives and Significance	70
4.4.2	Literature Review	71
4.4.3	Methodology	73
4.4.4	Experiments	80
4.4.5	Results and Discussion	82
4.4.6	Limitations	87
4.5	Providing Longitudinal Risk Predictions in Partial Annotation Scenarios	87
4.5.1	Objectives and Significance	87
4.5.2	Literature Review	89
4.5.3	Methodology	90
4.5.4	Experiments	96
4.5.5	Results	97

4.5.6	Discussion	101
4.5.7	Limitations	102
4.6	Integration and Discussion	103
4.7	Final Remarks	104
5	Robustness in ECG Diagnosis: Evaluating Multimodal Data Fusion Techniques	106
5.1	Introduction	106
5.2	Fundamentals of Multimodality	107
5.2.1	Definitions	107
5.2.2	Technical Challenges in Multimodal AI	108
5.3	Fusion Techniques	109
5.4	Improving Cardiovascular Disease Diagnoses using Electrocardiograms	111
5.4.1	Objectives and Significance	111
5.4.2	Literature Review	112
5.4.3	Methodology	115
5.4.4	Experiments	121
5.4.5	Results	121
5.4.6	Discussion	124
5.5	Final Remarks	127
6	Robustness in Continuous Patient Monitoring: Leveraging Asynchronous Inputs and Transformers	129
6.1	Introduction	129
6.2	Fundamentals of Longitudinal Data	130
6.2.1	Challenges	131
6.2.2	Longitudinal Modeling	131
6.3	Transformer Models	132
6.4	Developing Transformer Models for Critical Care	133
6.4.1	Objectives and Significance	133
6.4.2	Literature Review	135
6.4.3	Methodology	137
6.4.4	Experiments	148
6.4.5	Implementation Details	150
6.4.6	Results	152
6.4.7	Discussion	156
6.5	Final Remarks	157
7	Conclusion	159
7.1	Limitations and Future Directions	161
7.2	Contributions	163
7.2.1	Journal Papers	163

7.2.2	Conference Proceedings	165
7.3	Collaborations	166
7.4	Personal Reflections on AI and/or Tomorrow	167
Bibliography		169
Appendices		
A	Supplementary Material on Robust Cardiothoracic Surgery Risk Estimation	205
A.1	Inpatient Dataset Description	205
A.2	Detailed Experiments on Developing Risk Models for Cardiothoracic Surgery	207
A.3	Complementary Material on Handling Missing Data from the Real World	208
A.4	Detailed Experiments on Handling Missing Data from the Real World .	209
A.5	Detailed Experiments on Providing Longitudinal Risk Predictions in Partial Annotation Scenarios	212
A.6	Extended Results on Providing Longitudinal Risk Predictions in Partial Annotation Scenarios	213
B	Supplementary Material on Robust ECG Diagnosis	216
B.1	Detailed Experiments on Improving Cardiovascular Disease Diagnoses using Electrocardiograms	216
B.2	Extended Results on Improving Cardiovascular Disease Diagnoses using Electrocardiograms	219
C	Supplementary Material on Robust Continuous Patient Monitoring	221
C.1	Additional Data and Task Details on Developing Transformer Models for Critical Care	221
C.2	Detailed Experiments on Developing Transformer Models for Critical Care	225
C.3	Extended Results on Developing Transformer Models for Critical Care .	226

LIST OF FIGURES

1.1	Overview of this thesis structure.	8
2.1	Scheme of a CNN network.	15
2.2	Scheme of an RNN network.	15
2.3	Schematics of scaled dot product attention and MHA.	17
2.4	Schematic representation of the missingness mechanisms in the routine medical checkup example.	19
2.5	Example of an ML pipeline.	21
3.1	Levels of automation in medical AI systems.	34
3.2	Core domains to promote adoption of clinical DSSs.	35
3.3	Types of medical data at organizational and structural levels.	36
3.4	Proposed holistic approach to robust AI.	39
3.5	Ethical principles and key requirements for Trustworthy AI.	42
3.6	Risk levels in the AI Act.	44
3.7	Overview of Chapter 3 contributions.	47
4.1	Scheme of patient journey.	51
4.2	Training pipeline proposed for cardiothoracic surgery risk prediction.	61
4.3	Iterative feature selection performance evolution with new features.	63
4.4	Feature set associated with the best-performing model during the optimization phase.	63
4.5	Example of Shapley values obtained for each feature.	67
4.6	Depiction of the two types of correlation computed.	76
4.7	Evaluation pipelines of the missing data imputation in different datasets.	81
4.8	Depiction of the two types of correlation computed in CORKI	94
4.9	Overview of Chapter 4 contributions.	105
5.1	Scheme of the proposed ECG signal classification approaches.	112
5.2	Overview of the proposed 1D network for ECG classification.	117
5.3	Overview of the proposed CNN-based framework for ECG classification.	118

5.4	Overview of the proposed late fusion framework for ECG classification. . .	119
5.5	Overview of the proposed early fusion framework for ECG classification. . .	120
5.6	Overview of the proposed joint fusion framework for ECG classification. . .	120
5.7	Overview of Chapter 5 contributions.	127
6.1	Example of an ICU patient journey and generated data.	134
6.2	ICU-BERT architecture scheme.	145
6.3	Results for the ICU mortality classification task in terms of AUROC and AUPRC.	152
6.4	Results for the Kidney Function regression task in terms of MAE.	153
6.5	Results for the AKI Onset continuous classification task in terms of AUROC and AUPRC.	153
6.6	Overview of Chapter 6 contributions.	158
7.1	Overview of this thesis contributions.	164
A.1	Flowcharts of the CWKNNI and KNNSCI methods.	208
A.2	Flowchart of the CWRI method.	209
C.1	Few-shot results for the Hospital Mortality and Phenotyping classification tasks.	227

LIST OF TABLES

2.1	Example of missing data present in SBP measurements in routine medical checkups.	18
4.1	Inpatient dataset summary.	51
4.2	Outpatient dataset summary.	52
4.3	Registered complications in HSM.	53
4.4	Dataset summary for each population group.	59
4.5	Summary of the train, test A, test B, and complete sets.	60
4.6	Pipeline optimization top performances achieved through each model and feature importance method.	62
4.7	Final results validated against the test sets.	64
4.8	Results of weekly RPM kit distribution simulation.	65
4.9	Description of model performance during the clinical study.	68
4.10	Final results validated against the test sets.	69
4.11	Summary of missing rates across different variable groups in the cardiothoracic surgery dataset.	75
4.12	Results of missing imputation quality for all missingness mechanisms.	83
4.13	Results of classification performance for all classifiers.	85
4.14	Records, label, and missing values distribution in the multi-label cardiothoracic surgery dataset.	92
4.15	Records and label distribution in the CPSC, PTB-XL and G12EC datasets.	93
4.16	Results of ECG imputation for the test set in the CPSC, PTB-XL, and G12EC datasets.	98
4.17	Results of ECG classification for the test set in the CPSC, PTB-XL, and G12EC datasets.	99
4.18	Classification results of longitudinal risk estimation in the cardiothoracic surgery test sets.	100
5.1	Description of records per label in the PTB-XL dataset.	115
5.2	1D ECG classification results obtained with RNN and CNN-RNN models.	122

5.3	2D ECG classification results obtained with CNN models.	123
5.4	Multimodal fusion ECG classification results.	123
5.5	Final ECG classification results for best-performing models.	125
6.1	Summary of total samples in included datasets and fine-tuning tasks.	150
6.2	Performance of external evaluation and fine-tuning on the MIMIC-IV and YAIB datasets.	154
6.3	Performance of ICU-BERT in the YAIB tasks after fine-tuning on MIMIC-IV.	155
6.4	Performance of ICU-BERT model in the tasks replicated from DuETT after fine-tuning on MIMIC-IV.	156
A.1	List of variables by group in the inpatient dataset.	206
A.2	Hyperparameter ranges for the risk prediction model.	207
A.3	Tested parameter values for every imputation method within the comparative study.	210
A.4	Hyperparameter ranges for every classifier and dataset source.	210
A.5	Best imputation and classification hyperparameters and parameters for each dataset.	211
A.6	Hyperparameter ranges for every classifier and dataset source tested in CORKI.	212
A.7	Complete results of ECG classification for the test set in the CPSC, PTB-XL, and G12EC datasets.	213
A.8	Complete classification results of longitudinal risk estimation in the cardiothoracic surgery test sets.	215
B.1	Hyperparameter ranges for the unimodal and multimodal fusion models.	217
B.2	Optimal hyperparameters for each unimodal and multimodal fusion model.	218
B.3	Extended 1D ECG classification results obtained with RNN and CNN-RNN models.	219
B.4	Extended 2D ECG classification results obtained with CNN models.	219
B.5	Extended multimodal fusion ECG classification results.	220
C.1	Tables extracted from MIMIC-IV database and used in ICU-BERT.	222
C.2	Clinical concepts used by <i>ricu</i> to produce YAIB fine-tuning benchmarks.	222
C.3	ICU phenotypes used in the phenotyping fine-tuning task.	224
C.4	Hyperparameter ranges and selected values for ICU-BERT pre-training.	225
C.5	Hyperparameter ranges and selected values for ICU-BERT fine-tuning.	226

ACRONYMS

AGI	Artificial General Intelligence (<i>pp.</i> 45, 46)
AI	Artificial Intelligence (<i>pp.</i> 2–6, 8–10, 17, 21, 24, 26–34, 38–47, 49, 50, 53, 55, 103–109, 115, 124, 125, 127–129, 131, 159–163, 166–168)
AKI	Acute Kidney Injury (<i>pp.</i> 135, 149, 152–155)
APACHE	Acute Physiology and Chronic Health Evaluation (<i>p.</i> 140)
AUPRC	Area Under the Precision-Recall Curve (<i>pp.</i> 25, 151–156, 227)
AUROC	Area Under the Receiver Operating Characteristic Curve (<i>pp.</i> 25, 56, 57, 60–66, 69, 81, 84–86, 97, 99–101, 113, 151–156, 211, 213, 227)
BERT	Bidirectional Encoder Representations from Transformers (<i>pp.</i> 132, 133, 135–137, 146, 147, 150, 151, 155, 157)
BiGRU	Bidirectional Gated Recurrent Unit (<i>pp.</i> 117, 122, 218, 219)
BiLSTM	Bidirectional Long Short-Term Memory (<i>pp.</i> 117, 122, 218, 219)
BVP	Blood Volume Pulse (<i>p.</i> 114)
CABG	Coronary Artery Bypass Graft (<i>pp.</i> 56, 58)
CCMVI	Class Center-based Missing Value Imputation (<i>p.</i> 71)
CD	Conduction Disturbance (<i>pp.</i> 115, 121)
CMIM	Correlation Maximization-based Imputation Methods (<i>pp.</i> 72, 80, 83–85, 210)
CNN	Convolutional Neural Network (<i>pp.</i> 14, 15, 30, 113, 114, 117, 118, 122, 123, 217–219)
CNNI	Convolutional Neural Network Imputation (<i>p.</i> 72)
CNS	Central Nervous System (<i>p.</i> 59)
CoHiKNN	Correlation-based Hierarchical K-Nearest Neighbors (<i>pp.</i> 72, 76, 80, 83–85, 91, 210)
CORKI	Correlation-Optimized and Robust K Nearest Neighbors Imputation for Multi-Label Classification (<i>pp.</i> xii, 89–103, 160, 212, 213)
CPB	Cardiopulmonary Bypass (<i>p.</i> 56)

CPSC	China Physiological Signal Challenge 2018 (pp. 92, 93, 96, 98–100, 213)
CT	Computed Tomography (pp. 30, 35, 53)
CV	Cross-Validation (pp. 23, 24, 57, 60–62, 80, 81, 91, 93, 99, 148, 151–153, 227)
CWKNNI	Correlation Weighted K-Nearest Neighbors Imputation (pp. 76–79, 82–85, 103, 160, 208–211)
CWRI	Correlation Weighted Regression Imputation (pp. 79, 80, 83–85, 87, 103, 160, 208–211)
DL	Deep Learning (pp. 10, 13, 14, 16, 23, 27, 30, 73, 104, 106, 107, 110–117, 124, 125, 127, 131, 135)
DNA	Deoxyribonucleic Acid (p. 33)
DSS	Decision Support System (pp. 2–6, 8–10, 17, 21, 26–41, 45–49, 53–55, 63–68, 70, 73, 74, 86–89, 102–108, 124, 126–135, 137, 138, 155–163, 166, 168)
DT	Decision Tree (pp. 12, 60, 62, 96, 99, 207, 212)
EACTS	European Association for Cardio-Thoracic Surgery (p. 51)
EC	European Commission (p. 42)
ECG	Electrocardiography (pp. 1, 2, 4, 7, 9, 15, 47, 89, 91, 92, 96–99, 101–104, 106, 107, 109–130, 159–162, 213, 216)
EDA	Electrodermal Activity (p. 114)
EHDS	European Health Data Space (pp. 37, 161)
EHR	Electronic Health Records (pp. 5, 10, 17, 22, 29–31, 33, 35, 36, 134, 138–140)
ETH	Swiss Federal Institute of Technology (p. 141)
EU	European Union (pp. 28, 37, 42, 44, 46)
EuroSCORE	European System for Cardiac Operative Risk Evaluation (pp. 30, 48, 54–57, 66)
FCMI	Feature Correlation based Missing Data Imputation (p. 72)
FDA	Food and Drug Administration (p. 46)
FhAICOS	Fraunhofer AICOS (p. 50)
FHIR	Fast Healthcare Interoperability Resources (p. 37)
G12EC	Georgia 12-lead ECG Challenge (pp. 92, 93, 96, 98, 99, 213)
GAF	Gramian Angular Field (pp. 112, 114, 115, 117, 118, 124)
GDPR	General Data Protection Regulation (pp. 38, 44, 141)
GPT	Generative Pre-trained Transformer (pp. 45, 132, 133)
GRU	Gated Recurrent Unit (pp. 15, 112, 113, 117, 122–126, 131, 136, 154, 155, 218, 219)
HIPAA	Health Insurance Portability and Accountability Act (pp. 38, 139–141)

HiRID	High Time Resolution ICU Dataset (pp. 137, 138, 141, 144, 150, 153, 154)
HIS	Health Information System (p. 138)
HSM	Hospital de Santa Marta - Unidade Local de Saúde São José (pp. 6, 49–51, 53, 55, 58, 65, 68, 74, 88, 91, 103, 104, 166)
HYP	Hypertrophy (pp. 115, 121)
ICD	International Classification of Diseases (pp. 37, 134, 135, 156)
ICU	Intensive Care Unit (pp. 1, 2, 7, 17, 30–32, 47, 67, 129–131, 134–142, 144–146, 149–157, 159–162, 221, 224)
IEEE	Institute of Electrical and Electronics Engineers (pp. 43, 165)
IMDRF	International Medical Device Regulators Forum (p. 45)
IoT	Internet-of-Things (pp. 50, 52)
ISO	International Organization for Standardization (p. 43)
KDIGO	Kidney Disease Improving Global Outcomes (pp. 149, 154)
KNN	K-Nearest Neighbors (pp. 12, 71, 72, 77, 78, 80, 83, 85, 89, 90, 94, 96, 99, 210, 212, 213)
KNNSCI	K-Nearest Neighbors Selected by Correlation Imputation (pp. 78, 79, 83–87, 90, 91, 103, 160, 208–211)
LightGBM	Light Gradient Boosting Machine (pp. 13, 60, 154, 155, 207)
LLM	Large Language Model (pp. 132, 137)
LR	Logistic Regression (pp. 55–57, 154, 155)
LSTM	Long Short-Term Memory (pp. 15, 112–114, 117, 122, 131, 136, 154–156, 218, 219)
MAE	Mean Absolute Error (pp. 25, 80, 82–84, 86, 147, 152–155)
MAMIL	Multimodal Multi-Instance Learning Neural Network (p. 114)
MAR	Missing at Random (pp. 18–20, 72, 75, 82–84, 91)
MCAR	Missing Completely at Random (pp. 18, 19, 71, 73, 75, 82–85, 93)
MDR	Medical Device Regulation (p. 46)
MEDS	Medical Event Data Standard (p. 142)
MHA	Multi-Head Attention (pp. 16, 17)
MI	Myocardial Infarction (pp. 115, 121, 126)
MICE	Multivariate Imputation by Chained Equations (pp. 71, 73, 80, 83–85, 210, 211)
MIMIC	Medical Information Mart for Intensive Care (pp. 137–142, 144, 148, 151–157, 221, 222, 227)
ML	Machine Learning (pp. 3, 6, 8, 10–14, 21, 22, 24, 26, 27, 29–31, 34, 35, 43, 46, 55–58, 61, 65, 66, 71, 73, 81, 84, 86, 89, 92, 102, 104, 106, 111, 112, 115, 132, 134, 135, 209)

MLM	Masked Language Modeling (<i>pp. 133, 136, 147, 157</i>)
MLP	Multi-Layer Perceptron (<i>pp. 14, 96, 99, 100, 212, 213</i>)
MLVM	Masked Language-Value Modeling (<i>pp. 147, 225</i>)
MNAR	Missing not at Random (<i>pp. 18–20, 75, 83–86, 91</i>)
MRI	Magnetic Resonance Imaging (<i>pp. 35, 53</i>)
MSE	Mean Squared Error (<i>pp. 25, 26</i>)
MTF	Markov Transition Field (<i>pp. 112, 114, 115, 117, 118, 124</i>)
NB	Naïve Bayes (<i>pp. 13, 56, 60, 63, 81, 84, 85, 207, 209–211</i>)
NLP	Natural Language Processing (<i>pp. 15, 16, 30, 35, 41, 88, 128, 130</i>)
NMS	NOVA Medical School (<i>p. 50</i>)
NMVI	Nullify the Missing Values before Imputation (<i>pp. 72, 80, 83, 85, 210, 211</i>)
NN	Neural Network (<i>pp. 14, 40, 57, 110</i>)
NSP	Next Sequence Prediction (<i>pp. 133, 157</i>)
OECD	Organisation for Economic Co-operation and Development (<i>pp. 43, 44</i>)
PIC	Pediatric Intensive Care (<i>p. 138</i>)
PTB-XL	Physikalisch Technische Bundesanstalt XL (<i>pp. 92, 93, 96–100, 111, 113, 115, 116, 121, 126, 213</i>)
ResNet	Residual Neural Network (<i>p. 113</i>)
RF	Random Forest (<i>pp. 12, 24, 56, 60–63, 65, 81, 84–86, 96, 99, 207, 209–212</i>)
RMSE	Root Mean Squared Error (<i>p. 26</i>)
RNN	Recurrent Neural Network (<i>pp. 15, 16, 112, 113, 117, 119, 122, 131, 132, 136, 217</i>)
ROC	Receiver Operating Characteristics (<i>pp. 24, 25</i>)
RP	Recurrence Plot (<i>pp. 112, 114, 115, 117, 118, 124</i>)
RPM	Remote Patient Monitoring (<i>pp. 6, 49, 50, 53–55, 57, 61, 64–70</i>)
RQ	Research Question (<i>pp. 4, 5, 8, 159, 163, 164</i>)
SaMD	Software as a Medical Device (<i>pp. 45, 46</i>)
SBP	Systolic Blood Pressure (<i>p. 18</i>)
SF-36	36-question Short Form Health Survey (<i>p. 52</i>)
SGD	Stochastic Gradient Descent (<i>p. 217</i>)
SHAP	SHapley Additive exPlanations (<i>p. 66</i>)
SMOTE	Synthetic Minority Over-sampling Technique (<i>p. 22</i>)
SOFA	Sequential Organ Failure Assessment (<i>p. 30</i>)
SPECT	Single Proton Emission Computed Tomography (<i>pp. 74, 82, 86, 211</i>)
STFT	Short Time Fourier Transform (<i>pp. 113, 114</i>)

STS	Society of Thoracic Surgeons (<i>pp. 49, 54–57, 66</i>)
STTC	ST/T change (<i>pp. 115, 121, 126</i>)
SVM	Support Vector Machine (<i>pp. 13, 56, 60, 63, 81, 84, 85, 114, 209–211</i>)
TCN	Temporal Convolutional Network (<i>pp. 154, 155</i>)
TSFEL	Time Series Feature Extraction Library (<i>pp. 22, 106</i>)
UCI	University of California Irvine (<i>pp. 59, 74, 75, 80–82, 86, 209, 210</i>)
UK	United Kingdom (<i>p. 57</i>)
USA	United States of America (<i>pp. 38, 43, 46, 55, 56, 140, 166, 167</i>)
VOH.CoLAB	Value for Health CoLAB (<i>pp. 50, 166</i>)
WHO	World Health Organization (<i>p. 43</i>)
XAI	explainable AI (<i>p. 40</i>)
XGBoost	Extreme Gradient Boosting Machine (<i>pp. 56, 57, 156</i>)
YAIB	Yet Another ICU Benchmark (<i>pp. 135–137, 142, 144, 148–150, 152–155, 221, 222</i>)

PROLOGUE



Leo is a patient who, like many others, has a lifelong history of interactions with healthcare systems. In his early years, Leo visited his pediatrician for routine checkups, vaccinations, and occasional treatments for minor illnesses. When he went to school and participated in activities, a few careless moments led him to the hospital for a broken arm and a foot strain, among other minor injuries. As he grew older, Leo faced more complex health issues. At age 45, after a routine checkup, his doctor discovered elevated cholesterol levels and recommended lifestyle changes. Ten years later, Leo was diagnosed with hypertension using an Electrocardiography (ECG) exam, and after a series of tests, his doctor prescribed medication to control his blood pressure. Given his condition, Leo was referred to a cardiologist, who closely monitored him for early signs of cardiovascular disease. After several years of monitoring, the cardiologist recommended a valve replacement surgery to prevent worsening of the heart function. Over the years, Leo's medical history became rich with laboratory results, clinical notes, medication records, imaging data from chest X-rays, and even a genetic test to assess his risk for hereditary conditions.

Leo's journey through the healthcare system generated enormous data spread across multiple healthcare providers and systems. His medical records included both structured and unstructured historical data. The diverse data types are stored in different formats and systems that often do not communicate with each other. When Leo suffers a heart attack when recovering from the open-heart surgery, he is rushed to the hospital, where clinicians urgently need to access his medical history and determine the optimal care strategy. His condition requires not only rapid assessment but also continuous monitoring in the Intensive Care Unit (ICU) to prevent further complications. However, due to fragmented records, his care team struggles to retrieve all relevant information quickly.

This scenario illustrates one of the critical challenges in healthcare today: the complexity and lack of integration of medical data. Medical data has been collected and stored as a consequence of digitalization, but even when it is available, it's often incomplete, sparse, or difficult to interpret. As healthcare providers increasingly rely on digital health records, the sheer volume and diversity of data can overwhelm clinicians, leading to missed diagnoses, delayed treatments, and suboptimal care.

There is growing interest in leveraging Artificial Intelligence (AI)-based clinical Decision Support Systems (DSSs) to assist clinicians in making informed decisions. If designed effectively, these tools could help streamline care by analyzing large datasets in real time, identifying patterns, and making personalized recommendations for each patient. They hold the potential to improve diagnostic accuracy, guide treatment decisions, and enhance patient outcomes. For Leo, such systems could have enabled an earlier diagnosis of hypertension, identified his elevated risk for complications from cardiothoracic surgery, and provided real-time monitoring during his ICU stay, equipping clinicians with early warnings of deterioration and guiding critical decisions on when to perform life-saving interventions. However, the effectiveness of these systems depends heavily on the quality and robustness of the AI models and, ultimately, the underlying data.

The challenges of medical data complexity are compounded by the risks associated with untrustworthy AI models. Imagine if a DSS, rather than helping Leo's care team with reliable insights, relied on outdated or biased algorithms. An untrustworthy system might recommend the wrong medication based on flawed data or exclude critical factors, like his genetic predisposition to heart disease or prior surgeries. The consequences of such errors can be life-threatening, particularly in high-stakes situations like emergency care. Suppose AI models are not designed with robustness, transparency, and ethical considerations in mind. In that case, they risk losing the trust of clinicians and patients, ultimately undermining their potential to improve care in critical settings like cardiothoracic surgery, ECG diagnostics, and ICU monitoring.

INTRODUCTION

1.1 Motivation

The development of clinical DSSs holds great promise for transforming healthcare by enhancing diagnostic accuracy, optimizing treatment decisions, and improving patient outcomes. This potential has been amplified by the recent wave of products based on generative architectures, such as the ChatGPT [235], which have significantly reduced the gap between technology experts and the general population [125]. However, despite these advancements in AI and Machine Learning (ML) technologies, current systems face significant challenges that prevent them from reaching their full potential in real-world clinical practice. These challenges are primarily technical, but questions related to societal or environmental concerns also remain underexplored [125].

One of the most pressing issues lies in the complexity and heterogeneity of medical data [109]. As illustrated in the Prologue, data comes from various sources, including structured inputs like laboratory results and unstructured data such as clinical notes or medical images [216, 257]. These data streams often exist in silos, are collected at different times, and vary in quality and completeness [149]. Integrating this rich but sparse and asynchronous data into a coherent, reliable system poses a significant technical barrier. Moreover, the dynamic nature of healthcare, where patient conditions evolve rapidly, needs models capable of handling longitudinal data and providing timely, actionable insights [59].

Current clinical DSSs are often limited by narrow applicability, focusing on isolated tasks rather than addressing broader healthcare challenges. While they tend to perform well in controlled environments, these systems struggle in real-world settings, where missing data, diverse patient populations, and evolving clinical guidelines reveal deficiencies in robustness and reliability [303]. Additionally, many models do not align with clinical needs and lack transparency, leading to a decreased trust from clinicians, who may hesitate to rely on opaque algorithms for critical decision-making [43, 252].

In response to these challenges, this thesis proposes a holistic framework for developing robust AI for clinical decision-making. Our approach categorizes strategies for robust AI

into four core pillars: human-centric, data-centric, model-centric, and deployment-centric. Throughout this work, we explore three main lines of research that address these different strategies based on challenges posed by key use cases in clinical DSSs.

First, we adopt a clinician-in-the-loop approach in cardiothoracic surgery to develop risk models for predicting severe complications, addressing issues such as data heterogeneity and missing inputs. Second, we investigate the classification of cardiovascular diseases from ECGs, exploring how different representations of data in a multimodal framework can enhance robustness. Finally, we tackle the challenge of continuous monitoring in critical care, focusing on handling numerous and asynchronous medical variables.

Our overarching goal is to improve the robustness of AI-based clinical DSSs and increase trust among healthcare professionals, addressing the full spectrum of the AI system life cycle, from design to deployment.

1.2 Research Questions

In this thesis, we investigate the limitations and potential of DSSs within the healthcare domain, emphasizing the urgent need to improve robustness in the face of key challenges posed by real-world medical data. The following Research Questions (RQs) guide our efforts to develop more robust AI-based clinical DSSs:

RQ1. What are the key components of a comprehensive framework for developing robust AI in healthcare?

Real-world clinical environments present unique challenges that hinder the effective integration and adoption of AI-based DSSs. Ensuring robustness in AI systems used in healthcare requires a multidimensional approach that not only addresses technical factors but also incorporates human perspectives, data complexity, and operational constraints. This question seeks to identify the essential elements of a framework that can guide the development and deployment of robust AI in healthcare.

RQ2. What are the optimal strategies to address the challenges of heterogeneous, complex, and incomplete real-world medical data in clinical DSSs?

Medical data from real-world settings is often fragmented, incomplete, and context-dependent. Traditional approaches frequently overlook the nuances of clinical workflows and contextual biases. This question aims to find the most effective strategies for addressing these challenges by involving clinicians in data preparation. Key considerations include data processing methods, missing data handling, and optimal training pipelines to ensure robust models for accurate decision-making.

RQ3. How can clinical DSSs become more robust when dealing with multimodal data?

Healthcare data comes in various forms, including electrophysiological signals, imaging, clinical notes, and Electronic Health Records (EHR). Traditional systems often fail to fully utilize the richness of multimodal data, relying on a narrow set of representations. This research question explores how clinical DSSs can be more robust by designing models capable of integrating diverse data modalities to improve diagnostic accuracy and decision support.

RQ4. Do inferences made by clinical DSSs improve when leveraging complex and asynchronous historical information?

Longitudinal data is vital for tracking disease progression and predicting future health outcomes. However, real-world healthcare environments generate sparse, asynchronous, and irregular patient registries. Current clinical DSSs often simplify assumptions to incorporate this information partially. This question investigates whether leveraging the full complexity of asynchronous historical data can enhance the robustness of DSSs in dynamic and time-sensitive settings, such as critical care.

These research questions address the core challenges of developing and deploying robust AI-based clinical DSSs. Investigating clinician-assisted data preparation, multimodal integration, and longitudinal data usage aims to contribute valuable insight for developing systems that can operate effectively in real-world healthcare environments, ultimately leading to better patient outcomes.

1.3 Objectives and Contributions

The research questions identified and explored in this thesis translate into a set of objectives, each connected to specific medical use cases that present unique technical challenges and require distinct approaches. These objectives have led to meaningful contributions to the clinical DSSs field and align with the similarly numbered RQs. Below, each objective is outlined, along with its associated use case and contributions.

Objective 1: Analyze clinical DSSs limitations and design a holistic framework to improve robustness

This objective involves conducting a comprehensive review of the current clinical DSSs landscape to identify limitations and areas where AI can significantly enhance healthcare delivery. We start by evaluating the challenges posed by real-world medical data, such as heterogeneity, missing data, and lack of scalability, to propose a holistic framework that incorporates human-centric, data-centric, model-centric, and deployment-centric strategies. This framework serves as a guide for future research and development of

robust AI-based DSSs, addressing the entire system life cycle from data collection and preprocessing to deployment and post-deployment monitoring.

Contributions

- A holistic framework for robust AI in healthcare, integrating human-centric, data-centric, model-centric, and deployment-centric strategies.

Objective 2: Explore the potential of a clinician-in-the-loop approach to address the limitations of real-world heterogeneous medical data in cardiothoracic surgery risk prediction

This objective addresses the challenges of heterogeneous, real-world medical data in cardiothoracic surgery. In collaboration with clinicians from a Portuguese reference center, we propose a clinician-in-the-loop framework designed to enhance the quality of medical data, enabling the prediction of high risk for postoperative complications and improving patient selection for Remote Patient Monitoring (RPM). Leveraging over fifteen years of collected medical records, we also investigate correlation-based techniques to mitigate the impact of missing data and missing labels in longitudinal, multi-label risk models.

Use Case

Medical records from cardiothoracic surgery patients at Hospital de Santa Marta - Unidade Local de Saúde São José (HSM) encompassing clinical variables from pre-surgery to one-year post-procedure, based on extensive data collection and follow-up protocols implemented since 2008.

Contributions

- A clinician-in-the-loop pipeline for robust data processing.
- An optimization pipeline for training ML classification models development in complex scenarios.
- Deployment of a DSS in a clinical study to identify high-risk patients for remote monitoring.
- Supervised the development of three novel correlation-based missing imputation mechanisms for addressing missing data in the absence of expert inputs.
- A correlation-based label imputation technique to address missing labels in multi-label classification.

Objective 3: Study the potential of multimodal data representations for cardiovascular disease diagnosis

This objective explores the potential of multimodal fusion strategies to enhance robustness in multi-label classification tasks. Our focus is on using multiple data representations, such as raw ECG signals and signal-to-image transformations proposed in the literature, to improve the accuracy and reliability of cardiovascular disease diagnosis models.

Use Case

Multi-label classification of cardiovascular diseases using multi-lead ECG signals from a large public database, incorporating different data representations.

Contributions

- A comprehensive comparison of multimodal fusion strategies, evaluating the effectiveness of combining raw ECG data with signal-to-image transformations for cardiovascular disease classification.

Objective 4: Incorporate rich, asynchronous data for robust continuous monitoring in critical care

This objective focuses on understanding the complexities of critical care data, which are sparse, asynchronous, and come from multiple sources, including medical equipment and clinical registries. We aim to investigate the potential of attention-based transformer architectures to handle this data for continuous patient monitoring in the ICU. Within this objective, we also aim to identify key tasks in critical care that can verify the generalization capabilities of these models.

Use Case

Processing historical asynchronous data for continuous monitoring of patients in critical care, leveraging large ICU databases containing detailed medical registries and equipment-generated data.

Contributions

- A novel data structuring framework that organizes sparse, asynchronous, and rich medical variables without extensive preprocessing.
- A new transformer architecture designed to leverage the complexity of medical registries, overcoming limitations related to restricted variable sets and asynchronous data inputs.

By pursuing these objectives across different medical use cases, this thesis provides a holistic approach to tackling the challenges of developing robust AI-based clinical DSSs. The insights gained from these contributions form the foundation for developing and deploying DSSs that can operate with robustness in real-world clinical environments, improving patient outcomes and clinician trust in AI systems.

While this thesis aims to identify critical limitations and propose strategies for improving the robustness of clinical DSSs, it only covers some possible approaches. Further validation in real-world clinical settings, particularly across diverse patient populations and multiple clinical sites, is needed. Additionally, we did not explore other important robustness mechanisms, such as uncertainty estimation [24] and explainability techniques [103], which are crucial for building trustworthy AI systems. We advocate that achieving robustness in AI requires a holistic approach. Through this work, we propose a framework for addressing these challenges and investigate distinct strategies, laying the foundation for future advancements in this field.

1.4 Thesis Structure

This thesis is organized into seven chapters depicted in Figure 1.1, each addressing specific research questions, objectives, and use cases, ultimately contributing to the development of robust AI-based clinical DSSs in healthcare.

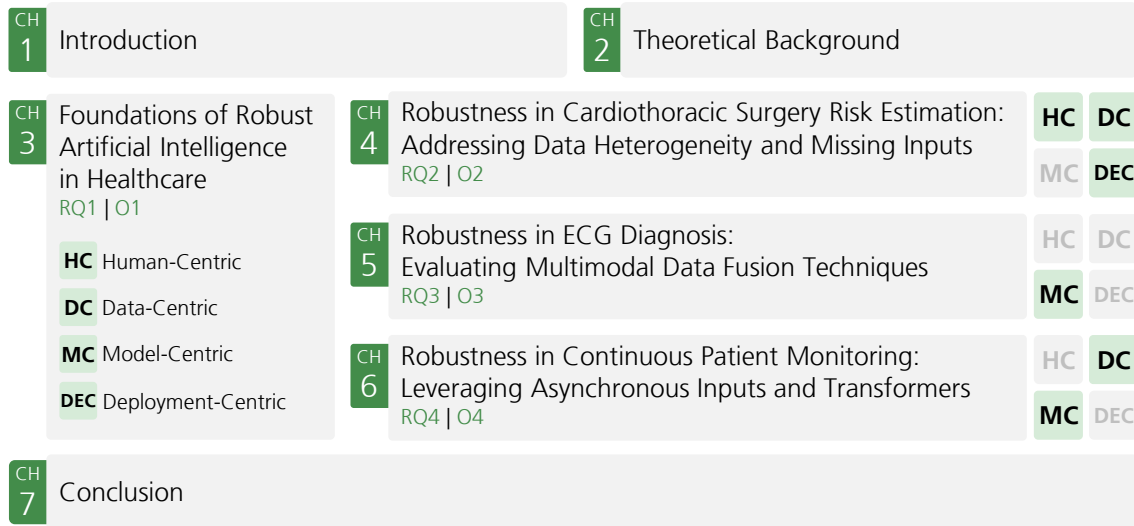


Figure 1.1: Overview of this thesis structure, detailing the alignment between chapters, RQs and objectives. Each technical chapter is aligned with specific approaches of the holistic framework proposed in Chapter 3.

Chapter 1 presented the context and motivation for the research, outlining the research questions, objectives, and key contributions of the thesis. Chapter 2 provides the essential theoretical background, introducing fundamental concepts related to medical data, as well as a primer on AI and ML, which serve as the foundation for the technical work in later

chapters. Chapter 3, aligned with [RQ 1](#) and [Objective 1](#), lays out the foundations for robust AI in healthcare, exploring both the opportunities and limitations of clinical DSSs and proposing a holistic strategy for robust AI development. In Chapter 4, corresponding to [RQ 2](#) and [Objective 2](#), we focus on the use case of cardiothoracic surgery, where we develop a clinical DSS and propose methodologies for handling missing data in real-world scenarios. Chapter 5 addresses [RQ 3](#) and [Objective 3](#) by investigating the potential of multimodal fusion in ECG classification, enhancing the robustness of disease detection models. The final technical chapter, Chapter 6, focuses on [RQ 4](#) and [Objective 4](#), exploring state-of-the-art strategies for processing longitudinal, asynchronous data in critical care, particularly for continuous monitoring. Finally, Chapter 7 summarizes the thesis contributions, discussing their impact on clinical DSSs and healthcare. It also outlines future research directions, emphasizing the importance of ongoing work, collaborations, and contributions to the scientific community.

THEORETICAL BACKGROUND

The potential of AI to enhance clinical DSSs is increasingly being recognized, offering significant opportunities to improve decision-making, diagnostic accuracy, and patient outcomes. However, deploying AI systems in real-world healthcare environments uncovers critical challenges, such as data heterogeneity, handling missing information, ensuring scalability, and adapting models to diverse clinical settings. Addressing these challenges is crucial for ensuring the robustness and effectiveness of AI-driven systems in healthcare.

This thesis focuses on improving the robustness of AI-based clinical DSSs. We aim to address the common issues associated with real-world healthcare data while ensuring the scalability and adaptability of the proposed ML and Deep Learning (DL) models.

The subsequent chapters present and discuss contributions to the scientific literature. However, providing a solid foundation of key concepts is essential to help the reader navigate this work. This chapter outlines the theoretical background necessary to understand the ML and DL techniques applied throughout the thesis. We introduce the fundamentals of ML and DL, including their characteristics and applications in healthcare. Next, we describe the crucial steps in processing medical data and the critical stages of the ML pipeline applied to healthcare settings.

2.1 Machine Learning

ML is a sub-field of AI that enables computers to learn patterns from data without being explicitly programmed for every task [294]. Instead of relying on hard-coded rules, ML models identify relationships in data and make decisions or predictions based on these patterns. This ability to "learn" and adapt makes ML an essential tool in many fields, especially healthcare, where data-driven insights can support clinical decision-making, predict patient outcomes, and personalize treatments.

In healthcare, DSSs based on ML models can help in diagnostic imaging by aiding radiologists in the detection of diseases such as cancer [203]; in forecasting hospital readmissions or disease progression using EHR data [291]; and in improving efficiency by extracting meaningful information from unstructured data, such as clinical notes and

patient records [134].

In the field of ML, numerous algorithms have been developed based on distinct mathematical foundations and paradigms, which can be categorized into various taxonomies. Regarding the learning procedure, the following types have been defined:

- **Supervised Learning:** Involves training a model to map inputs to outputs using labeled data [85]. This strategy aims to learn a function that predicts an output from a given input based on known input-output pairs. This task-driven approach is commonly used in classification and regression tasks. While in classification tasks, the objective is to separate the data, in regression, the models try to fit the data [281].
- **Unsupervised Learning:** Works with unlabeled data, identifying patterns and structures within the data without predefined labels or human intervention [108]. This data-driven method is used for clustering, anomaly detection, and dimensionality reduction, helping uncover hidden patterns and relationships [281].
- **Semi-Supervised Learning:** Semi-supervised learning is a hybrid of supervised and unsupervised learning [281]. It uses a small amount of labeled data combined with a larger set of unlabeled data to improve model accuracy. This method is appropriate for real-world scenarios where labeled data is scarce.
- **Reinforcement Learning:** Reinforcement learning trains models to make decisions by rewarding desirable behaviors and penalizing undesirable ones. The goal is to maximize cumulative rewards over time. It's commonly used in robotics, autonomous driving, and supply chain optimization [154].

The use cases and approaches studied in this thesis focus on supervised learning, adopting specific classification tasks with real-world relevance, such as predicting complications in cardiothoracic surgery, diagnosis of cardiovascular diseases, or identifying multiple outcomes in critical care.

Classification in ML tries to predict a class label for a given input. The goal is to learn a function $f : X \rightarrow Y$, where $X \in \mathbb{R}^n$ represents the input variables with n features, and Y represents the output labels or categories. Classification tasks are often categorized into binary, multiclass, and multi-label types [281]. Binary classification involves two class labels, i.e., $Y \in \{0, 1\}$, with one class typically representing a normal state ($Y = 0$) and the other an abnormal state ($Y = 1$), like "disease not detected" vs. "disease detected" in a medical test. For multiclass classification, the task involves mapping the input $X \in \mathbb{R}^n$ to one of several mutually exclusive class labels, i.e., $Y \in \{1, 2, \dots, k\}$, where k is the total number of classes. An example is classifying the severity of a disease, where each Y corresponds to a specific severity level. Multi-label classification is a more complex task where an example can belong to multiple classes simultaneously. This is represented as $Y \in \{0, 1\}^k$, where each of the k labels is a binary indicator, allowing multiple labels for

each example. For instance, a medical examination could be used by a model to predict several conditions for the same patient based on the same set of input features $X \in \mathbb{R}^n$.

2.1.1 Classification Models

ML models can be grouped into distinct categories based on their underlying learning strategies. When defining the learning objective, it is possible to foresee which types of models are more appropriate. Still, it is challenging to identify which model will provide the optimal results beforehand. In this thesis, we explored and tested different models that follow distinct learning paradigms.

Instance-based Learning

Instance-based learning models do not learn an abstract function during a training phase but base predictions on the similarity between new data points and stored examples [8].

K-Nearest Neighbors (KNN) is a simple, non-parametric algorithm that classifies a new data point by looking at the majority label of its nearest neighbors in the feature space according to some distance measure like the Euclidean distance. Although KNN is intuitive and easy to implement, it becomes computationally expensive as the dataset grows, making it less scalable for large clinical datasets.

Tree-based Learning

Tree-based models are a family of ML algorithms that rely on decision trees as their primary structure for making decisions or predictions [61]. These models partition the input data space into regions and then make predictions by assigning the majority class label or average value within each region. One of the key advantages of tree-based models is their interpretability based on the generated sets of rules.

Decision Trees (DTs) are the simplest tree-based models, where the algorithm recursively splits the data based on feature values to build a tree. Each split is chosen to maximize a measure of homogeneity (e.g., information gain or Gini impurity). Although very straightforward and a primary example of classification models, DTs are prone to overfitting, especially when the tree grows too deep [251]. Overfitting is a structural concern in developing ML models and consists of the excessive adaptation of the learning function to the training data, which harms the generalization to new cases that may only slightly deviate from the training distributions [78]. The opposite problem is underfitting, which, in contrast, leads to models with insufficient prediction capacity.

Ensemble methods combine multiple models to improve prediction accuracy and robustness. In the tree-based domain, Random Forest (RF) is an ensemble method that constructs multiple decision trees during training and aggregates their results to improve overall prediction accuracy and reduce overfitting, generalizing well even with limited data [48].

Building on the potential of tree-based models, gradient-boosting models iteratively create new trees to minimize prediction errors, with each correcting the mistakes made by the previous ones. Light Gradient Boosting Machine (LightGBM) is a specific implementation of gradient-boosting optimized for speed and memory efficiency, particularly with large datasets. It uses histogram-based learning and leaf-wise tree growth techniques to accelerate the learning process [159].

Probabilistic Learning

Probabilistic learning methods are based on probability theory, and predictions are made by modeling the underlying probability distributions of the data. These methods are particularly useful when dealing with uncertainty, as they quantify the likelihood of different outcomes. One of the most fundamental principles used in probabilistic learning is Bayes' Theorem, which provides a framework for updating the probability estimate for a hypothesis as more evidence or data becomes available [225].

The Naïve Bayes (NB) classifier is a simple but powerful probabilistic learning method that applies Bayes' Theorem with the strong assumption that features are conditionally independent given the class label [225]. Despite the assumption of independence often being unrealistic, NB performs well in many real-world applications. In its basic form, NB assumes that all features contribute independently to the probability of the class label. The NB can be generalized to use different distributions depending on the data characteristics, such as the Gaussian NB that assumes the continuous features follow a normal distribution, the Multinomial NB for discrete data; or the Bernoulli NB for binary-valued features [225].

Support Vector Machines

Support Vector Machines (SVMs) work by finding the optimal boundary, or hyperplane, that separates data points into different classes with the maximum possible margin [70]. They are beneficial in tasks where the decision boundary between classes is evident. The SVM algorithm is grounded in geometric principles and offers flexibility through its ability to work in high-dimensional spaces. While SVMs were initially designed for linearly separable data, real-world constraints led to the development of non-linear SVMs, which introduces the kernel trick to project the original data into a higher-dimensional space where it becomes linearly separable.

2.2 Deep Learning

DL represents a significant advancement over traditional ML, primarily driven by the exponential increase in computational power and the availability of vast amounts of data. While ML focuses on models that rely on handcrafted features to make predictions, DL automatically discovers and learns hierarchical representations from data, enabling

it to excel at tasks involving large, complex datasets, such as images, text, and audio [179]. This ability to autonomously extract meaningful features is one of the critical differences between ML and DL. In ML, algorithms require careful feature engineering, in which domain experts manually select and transform input features to make the model work effectively. Therefore, the quality of the handcrafted features limits the model's performance. On the other hand, DL models, particularly Neural Networks (NNs), can automatically learn features directly from raw data, eliminating the need for extensive manual intervention [114].

DL models are typically composed of numerous layers of interconnected neurons, which increasingly learn abstract representations of the input data as it passes through the network, allowing it to capture complex patterns and relationships [114]. In an image, early layers capture low-level features such as edges or textures, while deeper layers can learn high-level concepts like objects or scenes.

2.2.1 Deep Learning Strategies

With the increased relevance of DL, novel architectures have been proposed, which can be categorized into several types. Different architectures are optimal for specific scenarios and types of data.

Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) is one of the simplest forms of DL models and is often considered the foundation of NNs [114]. An MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer contains multiple neurons, and the connections between neurons are represented by weights adjusted during training.

MLPs are feed-forward neural networks, as information flows in one direction, from the input layer to the output layer, without any loops or cycles. They are most commonly used for tasks like classification and regression when the data is structured or tabular.

The training process of MLPs and other types of DL architectures is conducted through backpropagation. This algorithm iteratively adjusts the weights of connections in the network by propagating the error of the model's predictions backward through the network, updating the weights in each layer to minimize some loss function [338]. This process uses the gradient descent optimization technique, where the gradients of the loss are computed using the chain rule of calculus [114].

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are specialized for processing grid-like data, such as images and videos. An example of this network is depicted in Figure 2.1. The core idea behind CNNs is to exploit the spatial structure of data by using convolutional layers with optimizable filters that automatically detect local patterns, such as edges or textures

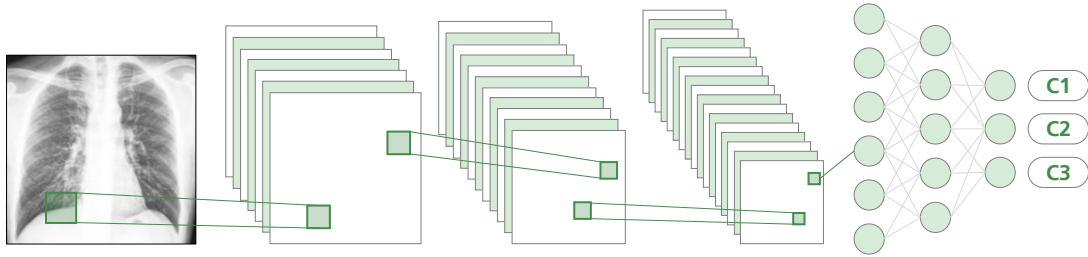


Figure 2.1: Example of a CNN network, featuring two convolutional layers followed by three feed-forward layers, leading to three output classes.

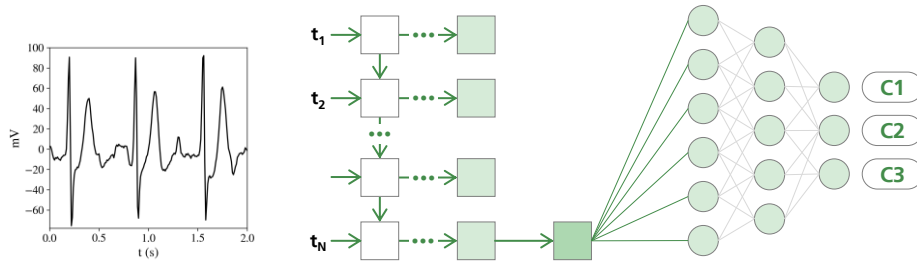


Figure 2.2: Example of an RNN network, with N inputs from an ECG signal, RNN layers followed by three feed-forward layers, leading to three output classes.

in images. These patterns are then combined across deeper layers to form more complex features, such as objects or faces [179].

CNNs have revolutionized computer vision, enabling tasks like image classification, object detection, and medical imaging analysis [16]. In addition to images, CNNs have been successfully applied to other domains like Natural Language Processing (NLP) and time series analysis by treating text or sequences as 1D grids [313].

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed to process sequential data, such as time series, speech, or text, as depicted in Figure 2.2. Unlike feed-forward networks, RNNs have loops that allow information to persist, making them ideal for tasks where the context or memory of previous inputs is essential [114].

Although showing potential in processing sequential data as memory of previous data points is kept, RNNs suffer from issues like vanishing gradients when dealing with long sequences. This led to the development of more advanced versions such as Long Short-Term Memory (LSTM) networks [116] and Gated Recurrent Units (GRUs) [67]. LSTMs and GRUs solve these problems by using gates to control the flow of information, enabling the model to retain relevant information for more extended periods.

RNNs and their variants have been widely applied in NLP, including tasks such as speech recognition, machine translation, and language modeling [315].

2.2.2 Attention Mechanism

The attention mechanism significantly advanced in DL, particularly for sequential and structured data tasks like language translation, speech recognition, and image captioning [230]. Unlike RNNs, where each element in the sequence is processed step-by-step and condensed into a fixed-size hidden state, attention mechanisms allow the model to focus on specific parts of the input data, regardless of their position in the sequence [326].

The key idea is to assign different "attention weights" to different input parts, indicating their relevance for making predictions. The attention mechanism has been widely adopted in models like transformers, which are now state-of-the-art in NLP tasks [76, 235].

At the core of the attention mechanism are three vectors derived from the input: Query (Q), Key (K), and Value (V). Each element in a sequence, the token, is transformed into these vectors, which serve distinct roles: Q represents the vector for which we want to find relevant information in the sequence; K represents the vectors that provide the information pertinent to the query; and V contains the actual information (context) that will be retrieved based on the relevance computed between the query and the key.

The attention mechanism computes a score to determine how much focus to place on each input part. This is done by taking the dot product of the query with all the keys, scaling by the dimension size of the key vector, and applying a softmax activation function to get the attention weights. The equation for scaled dot-product attention is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

In this equation, QK^T computes the dot product to obtain the similarity between the query and each key, and d_k is used as a scaling factor to prevent the dot product from growing too large in magnitude in cases of high-dimensional key vectors. The softmax is an activation function commonly used in DL to introduce non-linearity into models and allow learning complex relationships [231]. In this case, the softmax transforms the Q - K similarities into probabilities, which are then used to weigh the values V . This mechanism is defined as scaled dot product attention and is schematized in Figure 2.3a.

The attention mechanism is also known as self-attention when Q , K , and V are all derived from the same input sequence. In self-attention, the model computes attention scores across all tokens in the sequence, allowing each token to attend to every other token. This is particularly powerful in NLP tasks because it helps capture relationships between words regardless of their distance in the sequence.

In practice, models use h vector sets in parallel, known as Multi-Head Attention (MHA), as depicted in Figure 2.3b. Each attention head can focus on different input parts, making the model more expressive by learning different relationships between different tokens in parallel. The outputs from all attention heads are concatenated and linearly transformed to form the final prediction.

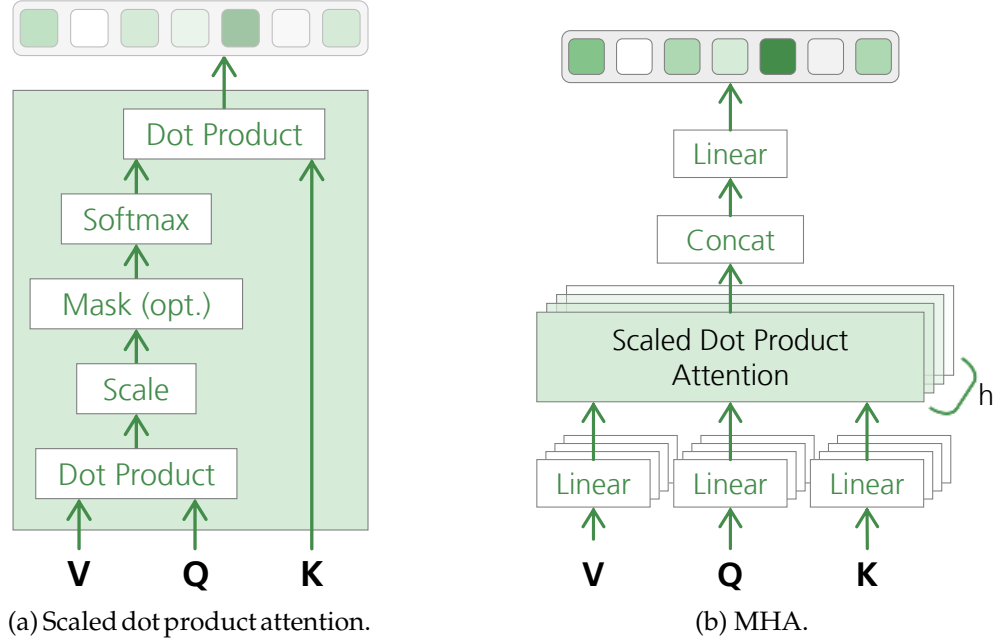


Figure 2.3: Schematics of scaled dot product attention and MHA. Q , K , and V are the queries, keys, and values, respectively, derived from the same input. h is the number of attention heads, and opt. meaning that the mask is optional. Based on [326].

2.3 Medical Data Processing

Medical data encompasses a broad range of information generated throughout a patient's journey within the healthcare system. This includes data from EHR, clinical notes, diagnostic imaging, physiological signals, laboratory test results, and prescription data. Medical data is the foundation for making informed medical decisions supported by AI-based clinical DSSs.

These diverse data sources can be categorized into different types based on their structure and nature [300]. Numerical data includes measurements like blood pressure, heart rate, and pulse rate, which can be further classified into discrete or continuous. Categorical data represents variables with distinct categories, such as disease types or medication groups, and can be classified as ordinal if values present a rank dependency or nominal when categories are independent. Binary data is a subset of categorical data where variables have only two possible values, such as "positive" or "negative" for a test result. Text data includes clinical notes, medical reports, and other unstructured documents. Additionally, time-series data is collected over time, often in ICU settings, where continuous monitoring of vital signs is common.

Combining these diverse data types makes medical data processing complex and valuable for building AI-based clinical DSSs.

2.3.1 Missing Data

Missing data is a common issue across various research domains, including healthcare. The presence of missing data in real-world datasets is ubiquitous and often overlooked in research works, although it can hinder the effectiveness of predictive models. Little and Rubin [191] define missing data as "unobserved values that would be meaningful for analysis if observed," emphasizing the loss of valuable information that, when left untreated, may lead to biased analyses or weakened statistical power. In medical contexts, missing data can result from several factors, such as faulty equipment, human errors in measurement or recording, or patient dropout during longitudinal studies.

To better understand and address missing data, Little and Rubin proposed a widely recognized classification system that distinguishes three mechanisms of missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR) [191]. These mechanisms describe how the missing data relates to the observed and unobserved values, which informs the selection of the most appropriate imputation strategy. Table 2.1 presents Systolic Blood Pressure (SBP) measurements registered during clinical checkups of different patients, together with their age. In addition to the complete registries, missing data originated by the distinct mechanisms is shown.

Table 2.1: Example of missing data present in SBP measurements in routine medical checkups, together with patients' age. MCAR, MAR, and MNAR refer to the missingness mechanisms applied to SBP. Retrieved from [71].

Age	SBP (mmHg)			
	Complete	MCAR	MAR	MNAR
11	110	110	?	?
16	126	?	?	126
21	108	108	?	?
29	133	133	?	133
35	96	96	96	?
40	100	?	100	?
43	152	?	152	152
56	116	116	116	116
67	148	148	148	148
71	140	?	140	140

In a MCAR mechanism, the missingness is unrelated to both the observed and unobserved data, meaning that the absence of data occurs purely by chance, without any systematic cause. For example, in a routine medical checkup, if a clinician forgets to record a patient's blood pressure or the measurement is omitted due to random technical issues, with no connection to the patient's health status or other variables, the missing data

would be considered MCAR. This randomness ensures that the missing data does not introduce bias into the analysis, although it can reduce the statistical power by decreasing the overall sample size. As shown in Table 2.1, the missing blood pressure values are distributed randomly across the different age groups, indicating no relationship between the missing values and other variables. Figure 2.4a visually illustrates this concept of random missingness, with a single association between the missingness and unmeasured variables.

In the MAR mechanism, the missingness depends on observed data but is unrelated to the actual missing values. For example, suppose clinicians are more likely to skip measuring blood pressure for younger patients because they are deemed at lower risk for hypertension. In that case, the missingness depends on the patient's age (an observed variable) but not the blood pressure value (the unobserved variable), as in Table 2.1. Figure 2.4b illustrates this scenario where the missingness is dependent on the observed age values, with also a potential relation to other unmeasured factors.

The MNAR mechanism occurs when the missingness is directly related to the unobserved data. In this case, the probability of missing data depends on the value of the missing data itself, making it much harder to address. For instance, if clinicians are less likely to record low blood pressure values because they consider them less clinically relevant, the missingness is related to the unobserved low blood pressure readings. This type of missingness introduces significant bias into the analysis because the absence of data is systematic, and appropriate correction methods are more complex. In Table 2.1, the missing blood pressure values are concentrated in the lower range of the blood pressure distribution, indicating a direct relationship between the missingness and the unobserved data. Figure 2.4c visually represents how MNAR missingness is tied to the unobserved low blood pressure values.

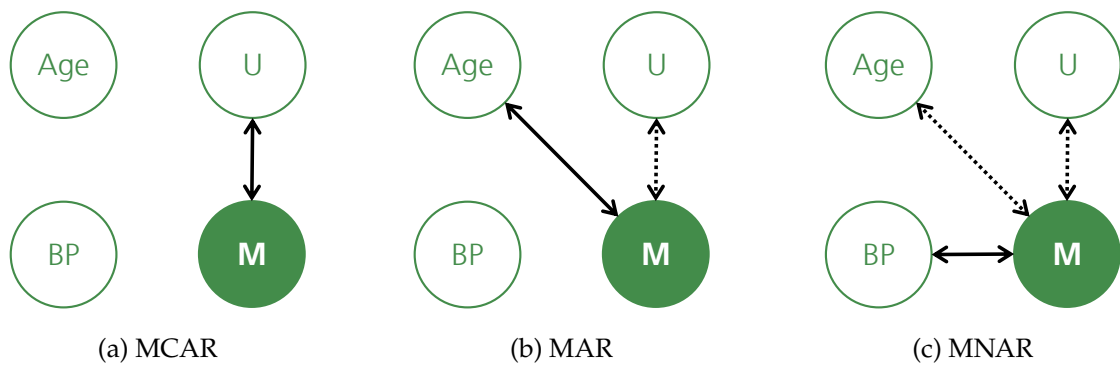


Figure 2.4: Schematic representation of the missingness mechanisms in the routine medical checkup example. The letters M and U represent, respectively, the missingness indicator matrix and a set of unmeasured variables. The solid arrows indicate statistical associations, while the dashed ones denote possible relationships. Based on [88].

Understanding these mechanisms of missingness is essential for selecting the most appropriate imputation strategy. MCAR data can often be handled by simple deletion

techniques without introducing bias, while MAR and MNAR require more sophisticated imputation methods to avoid skewing results.

2.3.2 Correlation

Correlation is a fundamental statistical tool that measures the degree of association between two variables. It is essential in medical research, where identifying relationships between variables can inform decisions about feature relevance and strategies for handling missing data. By assessing how changes in one variable are related to changes in another, correlation helps to predict trends and identify patterns [284]. Nevertheless, it is essential to note that correlation does not imply a causal relation.

Various correlation coefficients are used depending on the type of data being analyzed. The Pearson correlation coefficient, commonly used in healthcare studies, measures the strength of a linear relationship between two continuous, normally distributed variables. It is calculated using the formula:

$$r = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} \quad (2.2)$$

cov_{XY} is the covariance of variables X and Y , and σ_X and σ_Y are their standard deviations. The value of Pearson's coefficient ranges from -1 to +1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and +1 indicates a perfect positive correlation [12]. However, Pearson's correlation is only suitable for normally distributed numeric data and cannot be applied to other data types directly.

For binary variables, the Phi coefficient serves as an equivalent to Pearson's coefficient, measuring the correlation between two binary variables using the formula:

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad (2.3)$$

χ^2 is Pearson's chi-square statistic, and N is the number of observations. Like Pearson's, the Phi coefficient ranges between -1 and +1, reflecting the degree of linear association between binary variables.

The point biserial correlation coefficient is used when working with mixed data types, such as one binary and one continuous variable. It assesses the correlation between a binary variable Y and a continuous variable X and is given by:

$$r_{\text{pbi}} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_X} \sqrt{p_1 p_0} \quad (2.4)$$

$\bar{X}_1$ and $\bar{X}_0$ are the means of X for binary values $Y = 1$ and $Y = 0$, respectively. σ_X is the standard deviation of X , and p_1 and p_0 are the proportions of samples with $Y = 1$ and $Y = 0$, respectively.

Correlation has been studied in statistical analysis, and other methods have been widely used [71]. The presented correlation coefficients are based on Pearson's formula.

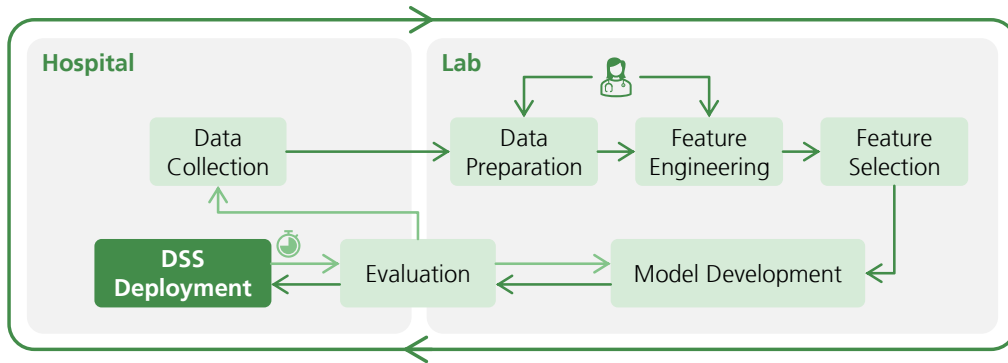


Figure 2.5: Example of an ML pipeline.

Therefore, the correlation values obtained from different variable types can be compared consistently across a dataset.

2.4 The Machine Learning Pipeline in Healthcare

The development of ML models goes through a series of well-defined stages before integrating a robust clinical DSS implemented in the real world. An ML pipeline is a structured process that transforms raw data into actionable insights [124]. An example of such a pipeline is schematized in Figure 2.5.

Numerous challenges demand serious considerations in healthcare, from the initial data collection to the deployment of predictive models. Due to the sensitive nature of medical data, standards and regulations enforce strict privacy and security protocols. Furthermore, the consequences of deploying AI models in clinical environments can be life-threatening. Incorrect predictions or recommendations could lead to wrong diagnoses or inappropriate treatments, potentially harming patients.

Healthcare providers and developers must collaborate closely to identify the clinical problem and align objectives with the needs of both clinicians and patients. Before developing any ML model, engaging stakeholders in defining the problem is essential. Understanding the intricacies of clinical workflows, patient care, and medical guidelines is critical to setting realistic goals and expectations. A literature review of similar applications of the problem can inform the development process and ensure that developed models are both scientifically grounded and practically applicable.

2.4.1 Data Collection and Preparation

Data collection in healthcare involves accessing, retrieving, and processing various types of medical data, often stored in separate systems. Establishing secure data retrieval pipelines is essential for connecting these systems. However, it is usually complex due to data heterogeneity, integration, and compartmentalization challenges, all discussed in Chapter 3.

The first step in data collection is setting up automated or manual pipelines to extract data from EHR, clinical notes, laboratory reports, imaging, and other medical sources while adhering to distinct standards and regulations specific to each healthcare context.

Anonymization of patient data is often a requirement before any further processing. Data masking, encryption, and de-identification are commonly used to ensure patient confidentiality [226]. Additionally, research ethics committees may be required to approve collecting and using healthcare data in scenarios involving interventions, ensuring compliance with ethical standards [86].

Once collected, the data is typically partitioned into training, testing, and validation sets. This step ensures that models are developed and evaluated using distinct datasets to avoid overfitting and to assess the model's generalizability. The existence of a validation set and the specific proportions depend on the characteristics of each use case, but the commonly used approach is an 80%/20% or 70%/15%/15% split. Most data is reserved for model training, with smaller portions for validation and testing.

Following data partitioning, several processing stages are required to ensure data quality [124]. Restructuring may be necessary if data comes from various sources and requires integration into a single format or dataset. Missing data is a common challenge in healthcare datasets. Depending on the nature of the missingness, techniques such as mean imputation, regression-based imputation, or more sophisticated algorithms may be applied. Detecting and handling outliers is crucial, as abnormal values in medical records could either reflect significant clinical abnormalities or be errors in recording. Normalization or standardization may also be applied to ensure that numerical data is on a consistent scale for gradient- and distance-based ML models.

In many healthcare applications, the collected data may suffer from class imbalance. For example, most instances may belong to the "healthy" category in a dataset used to predict rare diseases. Data augmentation or resampling (e.g., oversampling the minority class or undersampling the majority class) can address this imbalance. Synthetic data generation methods, such as Synthetic Minority Over-sampling Technique (SMOTE) [60], are also commonly used to rebalance datasets.

2.4.2 Feature Engineering and Selection

After completing data processing, the step in the ML pipeline that follows includes engineering new features and selecting the optimal set [281]. These stages are not always sequential and may require iteratively revisiting earlier phases.

Feature engineering involves extracting new features from raw data and transforming them into more accessible forms for ML models to interpret [172]. For example, categorical data can be transformed using one-hot encoding, converting several categories into a set of binary features. In healthcare, where longitudinal data like time-series signals are common, tools like Time Series Feature Extraction Library (TSFEL) [25] can be used to extract relevant features that capture temporal patterns, such as the frequency or variability

of signals. Additionally, combining features can reveal more predictive insights, as seen in calculating the body mass index from height and weight or deriving blood pressure gradients in cardiology.

Feature selection aims to identify the most relevant features for model training, reducing the dimensionality of the dataset while preserving predictive power [119]. One approach is correlation-based feature selection, where highly correlated features with the target variable but minimally correlated with each other are selected. Dimensionality reduction techniques like Principal Component Analysis [2] help condense high-dimensional datasets by projecting features into a lower-dimensional space while retaining most of the data's variance. Another strategy involves forward and backward selection methods, which either start with no features and progressively add them based on classification performance or start with all features and remove them until performance declines. Permutation feature importance is another effective method, shuffling the values of individual features to assess their impact on model performance, helping identify the most valuable predictors [15].

Feature engineering and selection can help reduce noise by removing non-informative features, enhancing model interpretability through smaller sets, and preventing overfitting. This is especially important in healthcare, where data is often high-dimensional and heterogeneous.

2.4.3 Model Development

The selection of the model candidates to train is essential, considering the different families of algorithms. This process is not always straightforward since the best-performing model for a given task is often unknown in advance and varies depending on the data and problem. Therefore, the typical strategy is to explore multiple algorithms, especially those that perform well in similar contexts. The selection process may involve choosing between the different categories presented in Section 2.1.1 or using DL techniques, depending on the complexity and size of the dataset and the problem. Extensive testing and evaluation through performance metrics are essential to identify the best candidate models. This experimental approach ensures that biases in model selection are minimized, as the ideal model is discovered empirically rather than assumed beforehand.

A critical aspect of model development is Cross-Validation (CV), a technique used to assess how well a model generalizes to unseen data. CV divides the dataset into multiple subsets or "folds," training the model on some folds and testing it on the remaining ones, with the process repeated several times [38]. There are various CV strategies to consider based on the specific structure of the data [47]. For example, K-fold CV partition the data into k folds, cycling through each as the test set. Stratified CV is commonly used in classification tasks, ensuring that each fold maintains the same proportion of class labels as the entire dataset. This is especially important in healthcare settings with imbalanced data. Group CV can be used when the data contains natural groupings or clusters, ensuring that

entire groups are kept together in the train/test splits. For more complex models, nested CV is employed, consisting of an outer loop that assesses generalization error and an inner loop for hyperparameter tuning, providing a more unbiased performance estimate.

During model training and combined with CV, hyperparameter tuning becomes essential to optimize performance [35]. Hyperparameters are configuration settings that influence model learning, such as the learning rate in gradient-based models or the number of trees in RFs. Grid Search is the most straightforward tuning method, where all possible combinations of a predefined set of hyperparameter values are exhaustively evaluated [35]. However, this can be computationally expensive, especially for high-dimensional search spaces. Random Search offers a more efficient alternative, where random combinations of hyperparameters are sampled and evaluated, often leading to a comparable performance with significantly fewer iterations [35]. More advanced techniques like Bayesian optimization take a probabilistic approach, building a surrogate model to estimate the performance of hyperparameter settings and using this information to make more informed choices in future iterations, thus speeding up the optimization process [298].

Careful consideration of model selection, CV techniques, and hyperparameter tuning allow for the development of robust models that generalize well to new, unseen data, especially in healthcare, where model reliability is critical.

2.4.4 Evaluation

The evaluation phase in the ML pipeline is critical for understanding how well a model generalizes to new, unseen data. Even when CV is used during model development, a separate test set is essential to provide a final, unbiased evaluation of the model's performance. This test set is kept entirely unseen during the training and tuning processes, acting as a proxy for how the model would perform in real-world scenarios. In healthcare, where decisions can significantly impact patient outcomes, reliable evaluation is crucial to ensure the model is safe, effective, and ready for deployment.

Different metrics can be used in combination to evaluate ML models, depending on the type of task and use case. In classification, most metrics work in the binary domain, which requires setting a decision threshold to transform probabilities into classes. While some works assume the 0.5 probability threshold, which is not ideal in imbalanced datasets, others try to optimize often using the Receiver Operating Characteristics (ROC) curve [245]. Among the existing metrics adopted by the AI community, we highlight the following [129]:

- **Accuracy:** Proportion of correct predictions out of the total predictions. While easy to interpret, it can be misleading for imbalanced datasets since the large ratio of negative samples versus positive tends to overestimate performance, as the models find negatives easier to identify.

- **Precision:** Measures the proportion of positive predictions that actually belong to positive samples. This is particularly useful in healthcare applications like disease diagnosis, where the dataset imbalance tends to return high rates of false positives.
- **Sensitivity (or recall):** Proportion of actual positive samples correctly predicted as positive classes. In medical contexts, it helps assess the model's ability to identify patients with a condition.
- **Balanced Accuracy:** The average of sensitivity across all classes. Balanced accuracy addresses the limitations of standard accuracy, especially in imbalanced datasets, by giving equal weight to the performance of each class. It can be used in binary or multi-class tasks.
- **Specificity:** Proportion of actual negative samples correctly predicted as negative classes, the opposite of sensitivity. In healthcare, specificity is important to determine how well a model identifies healthy patients with a negative class.
- **F1-Score:** The harmonic mean of precision and sensitivity, balancing both metrics, especially in cases of imbalanced data. High sensitivity accompanied by low precision denotes a model that tends to overestimate positive samples. In contrast, solely high precision may come with a model that avoids predicting positives, so the F1-Score acts as a balancing metric.
- **Area Under the Receiver Operating Characteristic Curve (AUROC):** Reflects the model's ability to distinguish between classes. AUROC is commonly used in binary classification tasks, as it measures how well the model distinguishes between classes across different thresholds. It is obtained from the ROC curve, a probability curve between the true positive rate (sensitivity) and the false positive rate (1 - specificity). This metric evaluates the models without fixing a decision threshold.
- **Area Under the Precision-Recall Curve (AUPRC):** More resilient to imbalanced datasets [275], provides a trade-off between precision and sensitivity. This metric emphasizes the model's performance in correctly predicting the minority class, summarizing how well the model identifies true positives while minimizing false positives across various decision thresholds.
- **Mean Absolute Error (MAE):** Utilized for regression, measures the average magnitude of errors in a set of predictions without considering their direction. This metric is intuitive and easy to interpret as it follows the original label unit.
- **Mean Squared Error (MSE):** Similar to MAE but squares the errors, making larger errors more penalizing. This can help identify models that produce large, unacceptable deviations.

- **Root Mean Squared Error (RMSE):** The square root of MSE provides an error metric in the same units as the target variable, making it easier to interpret.

Additionally, evaluating the model with external validation using real-world data that was not part of the training, tuning, or initial testing processes is often beneficial. This external evaluation helps assess the model's robustness when deployed in new environments, under different data distributions, or across various healthcare institutions. Models tested solely in laboratory conditions may not capture the full spectrum of real-world variability, so this step is crucial in healthcare applications.

2.4.5 Deployment

In the deployment phase of an ML pipeline, especially within healthcare, several important considerations must be addressed to ensure that the models deliver value in real-world settings while maintaining robustness.

Firstly, developing tools that align with users' needs is crucial. This includes designing intuitive user interfaces for healthcare professionals and integrating them into existing workflows without adding cognitive or operational burdens. For instance, clinicians must receive insights that enhance their decision-making process rather than complicate it, ensuring that the tool's outputs are interpretable and actionable.

Deploying an ML model in a solution that meets the needs is not the end of the process. Particularly in healthcare, the performance of a model must be evaluated regularly to ensure their predictions remain accurate over time [201]. Data distribution shifts can occur due to changes in patient populations, clinical practices, or new data collection methods, leading to potential degradation in model performance [106]. Continuous monitoring mechanisms allow real-time evaluation of a model's accuracy and reliability in its operational environment, ensuring it maintains robustness [341].

Following identifying performance degradation, mechanisms should be established to address the problem. ML models may need to be retrained with updated data [17]. Incorporating user feedback, updating models with recent data, and adjusting parameters for better performance are necessary steps to prevent model obsolescence. In a healthcare context, retraining models is critical as medical guidelines, diagnostic procedures, and patient demographics evolve.

Additionally, regulatory compliance and legal considerations must be addressed when deploying healthcare models. Ensuring the model's adherence to regulatory standards with secure and encrypted data pipelines and ethical usage is critical to maintaining trust and legality.

2.5 Final Remarks

Addressing the robustness of AI-based clinical DSSs in healthcare requires a solid theoretical foundation. While this thesis assumes that readers possess basic knowledge of ML,

this chapter has reviewed the most important concepts to support the work discussed throughout the thesis.

We began by outlining the fundamentals of ML and its various learning paradigms. Since the focus of this thesis is predominantly on classification tasks, especially in the development of DSSs for cardiothoracic surgery in Chapter 4, we examined different types of ML classifiers and the corresponding learning strategies they follow. This sets the preliminaries for understanding how ML models are employed in clinical settings.

Recognizing the limitations of traditional ML, the potential of DL is explored further in Chapters 5 and 6. To help the reader, we presented the key characteristics of DL and its principal network types, emphasizing the attention mechanism due to its significant contribution to recent advancements. As discussed in Chapter 6, the attention mechanism plays a crucial role in handling complex data patterns, offering distinct advantages over earlier methods.

Following the presentation of crucial AI concepts, we shifted our focus to essential considerations for processing medical data. These concepts, particularly relevant to the work in Chapter 4, include handling missing data and correlation measurements between different data types. We discussed the mechanisms behind missing data and the importance of understanding the relationships between variables, as these are fundamental to improving the robustness of clinical DSSs.

Finally, we discussed the crucial phases in developing a ML pipeline, from data collection and preparation to deploying models in real-world settings. These considerations, which are implemented in the technical chapters (4, 5, and 6), also form the basis of the holistic framework for robust AI proposed in Chapter 3.

FOUNDATIONS OF ROBUST ARTIFICIAL INTELLIGENCE IN HEALTHCARE

In recent years, significant advancements in AI research have revolutionized how society interacts with technology across various domains, with a notable focus on developing clinical DSSs. However, many AI models designed for healthcare applications often struggle to transition from the laboratory to real-world clinical practice. This gap is frequently due to flawed design processes that fail to adequately incorporate real-world inputs, an over-reliance on public datasets without rigorous external validation, and the absence of continuous supervision mechanisms [341]. These challenges underscore the need for a more robust approach to AI development in healthcare [53].

Enhancing the robustness of AI models for medical practice is not just a technical challenge but a fundamental objective that demands a multidisciplinary approach. The potential of AI in medicine is immense, offering opportunities to revolutionize disease diagnosis, treatment planning, patient monitoring, and personalized care [121]. However, this opportunity can only be fully realized by addressing the inherent limitations and risks associated with deploying AI in healthcare. These include concerns related to data quality, model interpretability, ethical considerations, and the reliability of AI systems in high-risk environments [109].

Recent efforts to develop guidelines and regulatory frameworks for trustworthy AI have highlighted the importance of carefully considering these limitations and risks. The European Union (EU) has been at the forefront of these efforts, establishing guidelines and the world's first comprehensive legal framework to ensure that AI technologies are developed and deployed in ways that respect fundamental rights, prevent harm, and are accountable and transparent [92, 130]. These initiatives are crucial for safeguarding human interests while promoting innovation in AI.

In this chapter, we will provide a comprehensive background that sets the stage for the specific challenges and methodologies explored in subsequent chapters. We will begin by discussing the fundamental principles that should underpin any medical AI initiative, including an exploration of DSSs in healthcare, their current impact, and the limitations

imposed by real-world data on their development. We will then categorize the approaches necessary to build robust AI models. Finally, we will examine the principles of responsible and trustworthy AI, discussing current guidelines, regulatory frameworks, and the ethical considerations that must guide the deployment of AI in healthcare.

3.1 Decision Support Systems in Healthcare

Clinical DSSs in healthcare are computer-based tools designed to assist healthcare professionals in making real-time, data-driven clinical decisions [37]. These systems leverage vast volumes of medical data to build a knowledge base that can be used to provide individual predictions during clinical practice. Common data types include demographic and socioeconomic variables, EHR, diagnostic tests, clinical notes, and other relevant medical data [147].

The primary purpose of DSSs is to enhance clinical decision-making by providing timely and accurate information that supports diagnostic, therapeutic, and patient management processes. Unlike traditional methods where clinicians rely solely on their expertise and available information, DSSs can analyze complex datasets, identify patterns or anomalies, and raise alarms or suggest actions that might not be immediately apparent [306]. These systems can range from simple applications offering basic reminders or alerts, such as drug interaction warnings, to more advanced systems that employ AI to predict patient outcomes [213], suggest treatment options [310], or assist in medical image interpretation [190].

Historically, DSSs have been implemented for various purposes, initially focusing on retrospective financial and administrative data analyses. Data mining approaches later expanded the scope of DSSs to include medical data for similar tasks [37]. However, it is important to distinguish between these early systems and clinical DSSs, designed to support real-time decision-making in clinical settings. Wright and Sittig [345] trace the origins of clinical DSSs to 1959 when Ledley and Lusted [180] proposed a probabilistic model for medical diagnosis based on set theory and Bayesian inference. This work provided clinicians with a framework for applying statistical and probabilistic reasoning to medical decision-making.

In 2007, Berner [36] proposed a taxonomy for clinical DSSs, distinguishing between knowledge-based and non-knowledge-based systems. Knowledge-based systems rely on predefined clinical rules and algorithms manually derived from clinical guidelines and expert input. In contrast, non-knowledge-based systems, which leverage AI and ML, learn from patterns in historical data rather than relying on predefined rules. These non-knowledge-based systems have become increasingly important for handling complex tasks and are now considered the most utilized tools in clinical DSSs.

The implementation of DSSs in clinical practice is not intended to replace the expertise of healthcare professionals but rather to augment their decision-making capabilities. By providing evidence-based recommendations, these systems help ensure that clinical

decisions are well-informed and consistent with best practices. Additionally, DSSs can improve efficiency within healthcare settings by automating routine tasks such as data entry, patient monitoring, and generating alerts for follow-up actions, allowing healthcare providers to focus more on direct patient care [217].

3.1.1 Current State

The current state of DSSs in healthcare is marked by the coexistence of traditional, statistically-based models and emerging AI-based tools that are gradually being integrated into clinical practice. Historically, clinical DSSs have primarily relied on statistical models to assist clinicians in making informed decisions. These systems typically employ scoring mechanisms or algorithms grounded in well-established clinical guidelines and evidence-based medicine.

For instance, in critical care, the Sequential Organ Failure Assessment (SOFA) score is widely used for predicting patient outcomes based on the degree of organ dysfunction [328]. The SOFA score is calculated using a combination of physiological measurements, laboratory data, and clinical parameters, providing an objective measure of illness severity that helps guide treatment decisions in the ICU. Similarly, in cardiac surgery, the European System for Cardiac Operative Risk Evaluation (EuroSCORE) is employed to estimate the risk of operative mortality [229]. This scoring system, derived from datasets of surgical outcomes, assists clinicians in assessing the risks associated with surgery and making informed decisions about patient management.

While these traditional models have been instrumental in standardizing care and improving outcomes worldwide, they are inherently limited by their reliance on predefined rules and static data. These models often fail to account for the complexity and variability of individual patient cases and do not leverage the vast amounts of real-time data generated in modern healthcare settings. This gap is where more advanced, AI-based clinical DSSs are beginning to make a significant impact.

Recent clinical DSSs leverage ML algorithms to analyze large and complex datasets, enabling the development of predictive models that can learn from data over time. These systems have shown promise in areas such as image segmentation, where DL models assist radiologists in detecting abnormalities in medical imaging. For example, CNNs have been applied to mammography and Computed Tomography (CT) scans, demonstrating high accuracy in identifying cancers and other conditions [90, 190]. Another significant area is text classification in EHR, where NLP techniques are employed to extract relevant clinical information from unstructured data, thereby aiding in diagnosis and treatment planning [134]. Moreover, recent attention-based models and multimodal transformers enable clinicians and patients to interact more intuitively with the system. This allows them to input tasks, like providing a mask or a text query, and receive accurate, task-specific outputs from the model. A notable example is Med-SAM [203], a medical image segmentation tool that operates in a zero-shot learning scheme, allowing users to provide a rough mask

around the region of interest without needing specific examples of segmentation targets. This tool holds enormous potential for improving medical image evaluation, although its validation and implementation in real-world practice are still pending.

Risk models, particularly those incorporating AI, are also gaining traction in clinical settings. These models predict the likelihood of various outcomes, such as hospital readmission [289] or the development of complications [223], by analyzing patterns in patient data. For instance, ML algorithms have been developed to predict the onset of sepsis in ICU patients, enabling earlier intervention and potentially saving lives [170]. Similarly, AI models have been used to predict hospital mortality [107], offering clinicians valuable insights into patient prognosis and helping to inform care decisions.

Despite these advancements, many AI-based clinical DSSs are still in the early stages of integration into clinical practice. Challenges such as data quality, model interpretability, and the need for rigorous validation across diverse clinical settings remain significant barriers to widespread adoption [109, 319]. Nevertheless, the ongoing development of these systems holds great promise for transforming healthcare, offering more personalized and precise medical interventions tailored to each patient's unique needs.

3.1.2 Limitations

The current landscape of clinical DSSs offers significant promise for enhancing healthcare delivery. However, several critical limitations hinder their widespread adoption and effectiveness in real-world clinical settings.

A primary challenge is the integration of clinical DSSs within existing healthcare information systems. Although healthcare institutions have been collecting and storing vast amounts of medical data, these systems often operate on an outdated infrastructure that lacks the advanced functionalities required by AI-based DSSs [7]. The healthcare systems of multiple countries, already facing rising costs that do not correlate with better patient outcomes [319], have limited capacity for the large-scale investments needed to upgrade these systems. Furthermore, the lack of interoperability between different EHR systems even within the same hospital exacerbates the problem, creating silos of patient information that are difficult to access and integrate [7, 149]. As a result, clinical DSSs often work with fragmented data, limiting their ability to provide accurate predictions and comprehensive recommendations.

Beyond data access issues, current clinical DSSs struggle with the inherent complexity of medical data. Medical data is diverse and originates from multiple sources, is collected at different times, and varies significantly between patients [249]. Moreover, it is inherently longitudinal, capturing a patient's health trajectory over time. Traditional DSSs are typically designed to process static snapshots of patient data, which restricts their ability to generate insights from the patient's entire medical history. Additionally, the lack of multimodal data integration further limits the efficacy of these systems.

Compounding these issues is the challenge of data quality [337]. Healthcare data

often contains errors, inconsistencies, and missing values [109]. Such issues can lead to inaccurate predictions, which could have severe consequences in a clinical setting. Variations in data quality across institutions due to differences in data entry practices, measurement techniques, and coding standards further complicate the deployment of reliable clinical DSSs.

Another significant barrier to the adoption of clinical DSSs is the acceptance by healthcare professionals. Clinicians may resist new technologies, particularly when they perceive them as complex, unreliable, or a risk for loss of autonomy [177]. This resistance is often due to the "black box" nature of AI-based systems [56]. In high-risk environments like ICUs, clinicians may hesitate to trust recommendations from a system that does not explain its outputs clearly. However, thorough research on the impact of explainable systems is yet to be studied. Cabitza et al. [54] coined the term "white box paradox," which refers to a risk of explainable systems being more effective in misleading humans in cases where they provide erroneous predictions. Moreover, poorly designed clinical DSSs can increase the cognitive workload of healthcare providers, contributing to "alert fatigue," where clinicians become overwhelmed by the system's alerts and begin to disregard them [161].

Scalability and generalization of clinical DSSs across different healthcare settings also present significant challenges [28]. A model that performs well in one hospital may not generalize to others due to differences in patient demographics, clinical practices, and data collection methods. This issue is particularly pronounced in global health contexts, where variations in resources and disease prevalence can impact the effectiveness of DSSs. Moreover, medical knowledge continually evolves, and many clinical DSSs are not equipped to incorporate new evidence or guidelines as they emerge. This lack of adaptability can lead to concept drift, where the model's performance degrades over time as the underlying data distribution changes [201].

Finally, the deployment of AI-based clinical DSSs raises ethical, social, and legal concerns, particularly around patient privacy, data security, and the potential for bias [120, 166]. AI models trained on biased datasets can inadvertently perpetuate or even exacerbate health disparities. Ensuring that clinical DSSs are developed and deployed in ways that are fair, transparent, and aligned with ethical standards is essential [163]. The legal landscape for AI in healthcare is also still evolving, with ongoing debates about liability and the regulation of AI-based medical devices [324].

While DSSs hold immense potential to transform healthcare, their widespread adoption is hampered by significant challenges, including integration within existing systems, handling the complexity and quality of medical data, achieving clinician acceptance, and ensuring scalability and ethical deployment. Addressing these challenges is crucial for developing robust, reliable, and effective clinical DSSs that can enhance patient outcomes in real-world settings.

3.1.3 Future Potential

As AI research advances, particularly in improving the robustness and accuracy of clinical DSSs, we can foresee a future where intelligent, data-driven technologies significantly enhance clinical decision-making.

One of the most promising developments in future clinical DSSs is the potential to create more efficient administrative and clinical processes [75]. AI-based systems could dramatically reduce the time and effort required for data analysis, diagnosis, and treatment planning. By automating routine tasks such as data entry and patient monitoring, healthcare providers could focus more on direct patient care, leading to quicker response times and better patient outcomes.

Another transformative area for AI-based DSSs is the early detection and forecasting of diseases [249]. The ability of AI to analyze vast datasets could enable future systems to predict the likelihood of a patient developing a particular condition years before it manifests. This would open a crucial window for preventive measures and early treatments, significantly improving patient prognosis. For example, AI models could analyze genetic data to identify individuals with high risk for cancers [358] or cardiovascular diseases [27, 209], allowing for personalized screening programs and early interventions tailored to each patient. Integrating genomics and DNA sequencing into DSSs could offer highly personalized treatment recommendations, such as predicting individual responses to treatments like chemotherapy [74]. This approach represents a shift from the one-size-fits-all model to truly personalized healthcare. Nevertheless, aiming at these personalized medicine opportunities brings additional challenges related to the potential for bias, which needs to be addressed [240].

Beyond disease prediction and treatment, AI-based DSSs could play a pivotal role in enhancing patient engagement [75]. By providing tailored health recommendations, monitoring progress, and offering real-time feedback, these systems could encourage more active participation in healthcare. This could be particularly beneficial for managing chronic conditions, where adherence to treatment plans is crucial, and extend to preventative health by offering personalized lifestyle recommendations based on continuous monitoring and data analysis [206].

One of the most ambitious visions for future clinical DSSs is the concept of a "digital twin": a comprehensive digital replica of a patient that integrates all data collected throughout their life, from birth to present [50]. This digital twin would continuously update with new information, including EHR, genetic data, lifestyle factors, and real-time health monitoring, providing a holistic view of the patient's health [157]. Such a model could simulate different treatment options, predict disease progression, and offer highly personalized healthcare recommendations [325]. While the potential for this technology is enormous, it also presents significant technical, ethical, and privacy challenges that must be addressed [305, 155].

A key concern that hinders clinician trust and acceptance of clinical DSSs is the fear of

	Assistive AI		Autonomous AI		
	Level 1 Data Presentation	Level 2 Decision Support	Level 3 Conditional Automation	Level 4 High Automation	Level 5 Full Automation
Monitoring	AI	AI	AI	AI	AI
Execution	Clinician	Clinician + AI	AI	AI	AI
Fallback	Not Applicable	Clinician	AI with Backup	AI	AI
Liability	Clinician	Clinician	Case Dependant	Developer	Developer

Figure 3.1: Levels of automation in medical AI systems. Adapted from [41].

losing autonomy in decision-making. Bitterman et al. [41] proposed a framework of five levels of automation, as depicted in Figure 3.1, that affect responsibility and liability: data presentation, clinical decision support, conditional automation, high automation, and full automation. While the latter stages involve minimal to no clinician oversight, they may not be achievable in high-risk environments like healthcare. Topol [319] argues that medicine will likely remain at level 3, where humans retain oversight over AI’s interpretation of images and data. The true potential lies in Human-AI collaboration, where AI can process and analyze data on a scale and speed beyond human capabilities. Yet, the ultimate decisions in healthcare remain with clinicians. AI can assist by uncovering patterns and insights that might be missed by human analysis, thus enhancing the clinician’s expertise and enabling more informed, precise decisions [287].

The realization of these future clinical DSSs capabilities will require continued advancements in AI and ML, but the room for improvement goes beyond technical details. Adler-Milstein et al. [7] propose a framework, schematized in Figure 3.2, that organizes the drivers for successful adoption of clinical DSSs in four domains: the *reason to use* domain requires aligning incentives and market forces to drive healthcare investment in AI-based DSSs; the *means to use* domain emphasizes the importance of robust data and human infrastructure, along with the necessary technical resources to deploy and maintain AI-based DSSs in clinical environments; the *method to use* domain highlights the need to consider workflow integration and the training requirements necessary for clinicians to use these tools effectively; and the *desire to use* domain addresses the psychological factors influencing clinicians comfort with AI, including how these tools can reduce burnout, provide professional fulfillment, and build trust.

3.2 Real World Medical Data

Medical data forms the backbone of clinical DSSs, providing the essential information needed to build the algorithms and models to guide healthcare professionals in improving patient care through informed, data-driven decisions [307]. This data, generated and collected throughout patients’ lives, comes in various forms and formats, each contributing

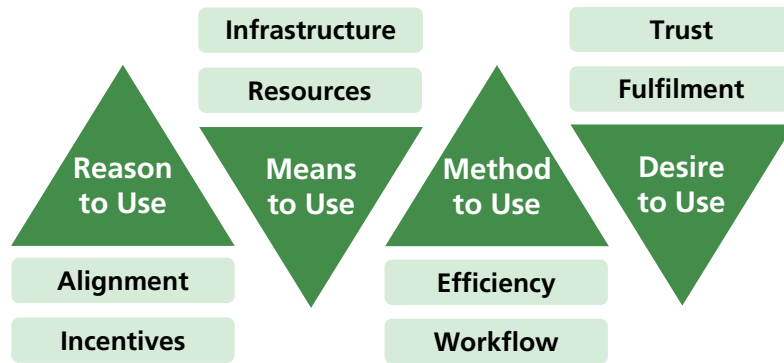


Figure 3.2: Core domains to promote adoption of clinical DSSs. Adapted from [7].

uniquely to a comprehensive understanding of patient health. The quality, availability, and structure of this medical data directly impact the effectiveness of clinical DSSs. Still, its diversity and complexity also introduce significant challenges in processing and integrating [109].

Medical data can be broadly categorized into three types based on its organization: structured, unstructured, and semi-structured. Structured data is highly organized, fitting neatly into tables or relational databases, and includes discrete fields like EHR. This data type is easily stored and searchable and can be directly used in traditional ML models [257]. Unstructured data, on the other hand, lacks a standardized structure, making it more challenging to search, process, and analyze [216]. This category includes free-text clinical notes, raw imaging data, and audio recordings. Extracting meaningful insights from unstructured data often requires sophisticated techniques such as NLP for text and image recognition algorithms for medical imaging. Semi-structured data lies somewhere between these two extremes; it retains some organizational structure, such as metadata in medical imaging files or genomic sequences, but still requires specialized processing to be fully utilized. While structured and semi-structured data originated from instrument readings and the digitization of health records, unstructured data is often generated at the point of care [257].

In addition to organizational categories, medical data varies in its structural format. Tabular data is typically structured and includes EHR, laboratory results, and demographic information. Textual data is unstructured and includes various sources such as clinical notes and exam reports. Image data, including X-rays, Magnetic Resonance Imaging (MRI), and CT scans, is complex. While the visual content is unstructured, image data is often accompanied by semi-structured metadata that provides additional context. Genomic data, which is also semi-structured, contains sequences along with associated metadata, and time-series data, which captures physiological signals over time, demands models capable of processing temporal dependencies. These data structures present unique challenges for clinical DSSs, particularly regarding data integration, model interpretability, and validation. Figure 3.3 summarizes the different types of medical data according to

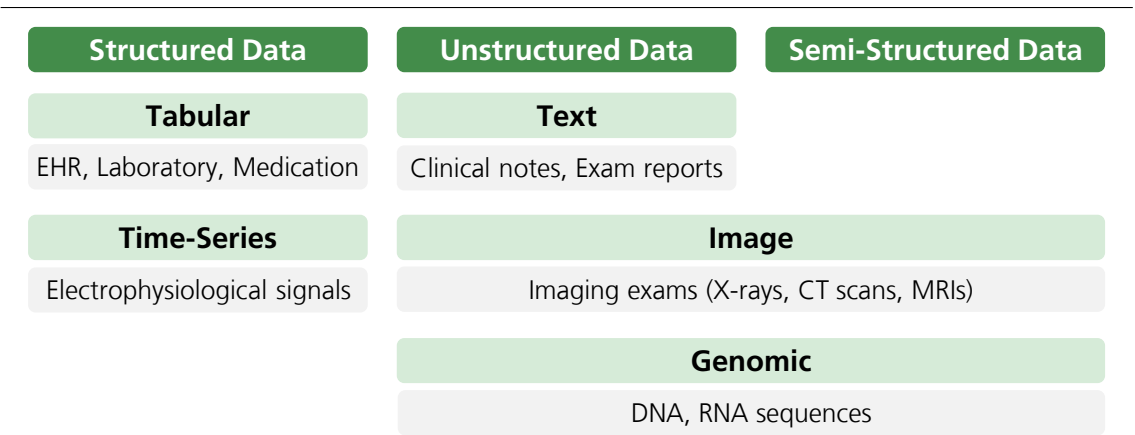


Figure 3.3: Types of medical data at organizational and structural levels.

organization and structure.

A crucial feature of medical data is its longitudinality, i.e., the ability to capture a patient’s health trajectory over time [59]. Different types of data offer varying degrees of longitudinality. For instance, time-series data and EHR are inherently longitudinal, tracking changes in health status in real-time or over extended periods. In contrast, imaging data typically provides only snapshots at specific moments, and genomic data generally represents a static profile. Longitudinality is essential for understanding disease progression and developing robust clinical DSSs that offer actionable insights throughout a patient’s healthcare journey [102].

Despite the richness and diversity of medical data, its variability and lack of uniform structure present significant challenges in integration and analysis. These challenges must be addressed to build robust clinical DSSs that improve patient outcomes across various healthcare settings.

3.2.1 Challenges

Leveraging real-world medical data to develop robust, reliable, and scalable clinical DSSs presents numerous challenges that contribute to the limitations of current DSSs, as discussed in Section 3.1.2. These challenges can be broadly categorized into compartmentalization, data quality, temporal and modal complexity, bias, privacy, and ethics.

Compartmentalization

Compartmentalization refers to the fragmentation of medical data across different systems and departments, making it difficult to access and integrate all relevant patient information [109]. Medical data is often siloed within various clinical EHR systems, each using its standards, formats, and protocols. This lack of interoperability creates barriers to data integration, preventing DSSs from accessing a comprehensive view of a patient’s health [149]. The compartmentalization of data not only hampers the development of effective models but also leads to inefficiencies in clinical workflows, as healthcare providers may

need to manually gather and reconcile data from disparate sources, resulting in wasted resources and repeated examinations. Despite efforts to standardize health data through initiatives like Fast Healthcare Interoperability Resources (FHIR) [330], many healthcare institutions still use legacy systems that were not designed to support the advanced functionalities required by modern DSSs [207]. This lack of harmonization often results in fragmented, incomplete data, limiting the potential of clinical DSSs to deliver accurate and actionable insights.

However, recent initiatives like the European Health Data Space (EHDS) [91] hold promise for the future by aiming to create a unified framework for accessing and sharing health data across EU member states, which could improve data interoperability and support the development of more robust, integrated clinical DSSs.

Data Quality

Medical data often suffers from errors, inconsistencies, and missing values, which can severely compromise the reliability of clinical DSSs. Errors can stem from data registration mistakes, variations in documentation practices across institutions, inherent noise artifacts, and measurement inaccuracies [149]. These issues contribute to data corruption, making it challenging to develop reliable models. For instance, laboratory results might be incorrectly labeled or misfiled, leading to flawed analyses and predictions. Additionally, the imprecision of outcome definitions, such as using International Classification of Diseases (ICD) codes primarily intended for billing, can lead to misleading labels for disease severity or mortality, deviating from the accurate representation of a patient's condition [268].

Furthermore, medical data is often incomplete, with missing values arising from various sources and subject to different mechanisms [191], such as skipped measurements or unrecorded patient interactions. These gaps can introduce significant bias into models if not appropriately handled, leading to erroneous predictions and potential harm in clinical settings [109].

Temporal and Multimodal Complexity

The inherent complexity of medical data is not only due to its diverse sources but also its heterogeneous nature. Different data types, such as imaging, lab results, and clinical notes, are collected at varying times and under different circumstances. This asynchronicity can create data gaps that are difficult to bridge, complicating analysis and modeling. Additionally, medical data is often longitudinal, capturing a patient's health trajectory over time [59]. The variability in data collection intervals and differing levels of granularity across data types (e.g., continuous vital signs vs. sporadic lab results) pose significant challenges for models attempting to accurately reflect patient health over time. While leveraging all data sources would ideally provide a holistic view of the patient, technical challenges often undermine the success of truly multimodal models. Moreover, not all

data modalities are always available, leading to situations where critical information might be missing or outdated, further complicating the integration of these diverse data sources into a coherent model [149].

Bias, Privacy, and Ethics

Using real-world medical data in clinical DSSs raises significant concerns related to bias, privacy, and ethics. Bias is a pervasive issue in medical data that poses substantial challenges for developing clinical DSSs. Bias can be introduced at multiple stages, from data collection and annotation to model training and deployment [341]. This issue can arise from various sources, including race, demographic disparities, socioeconomic factors, and historical inequalities in healthcare access [4, 233]. These biases can lead to models that perform well for specific populations but poorly for others, exacerbating existing health disparities. For example, a clinical DSS trained primarily on data from urban hospitals might not generalize well to rural populations, leading to less effective care for those patients. Ensuring fairness in clinical DSSs is crucial to avoid perpetuating or exacerbating health inequities that lead to suboptimal care for marginalized groups [237].

Privacy and ethical considerations are also important when dealing with medical data. Healthcare data is highly sensitive, containing personal information that, if mishandled, could lead to breaches of patient confidentiality. Ensuring the privacy and security of medical data is essential, particularly as healthcare systems become increasingly digital and interconnected [211]. Regulations like the Health Insurance Portability and Accountability Act (HIPAA) [1] in the United States of America (USA) and the General Data Protection Regulation (GDPR) [93] in Europe provide frameworks for protecting patient data. Still, they also introduce challenges for data sharing and use in clinical DSSs development. Balancing the need for robust, data-driven clinical support with the ethical obligations to protect patient privacy and prevent biased outcomes is a complex but essential aspect of advancing clinical DSSs technology.

3.3 A Holistic Approach to Robust AI

Robustness refers to the development of AI systems that can perform reliably under a wide range of conditions, including varying data quality, unforeseen inputs, and complex real-world environments [130]. Even when AI systems are designed with ethical principles at their core, it is crucial to ensure they do not cause unintended harm.

Robustness in AI involves a dual focus: technical robustness and robustness from a social perspective [130]. Technical robustness refers to the AI system's ability to perform consistently under different conditions, maintaining accuracy, reliability, and security across its life cycle. This includes resilience to errors, adaptability to new situations, and the capacity to handle unexpected inputs without compromising performance. In the context of clinical DSSs, technical robustness is essential for ensuring the system can

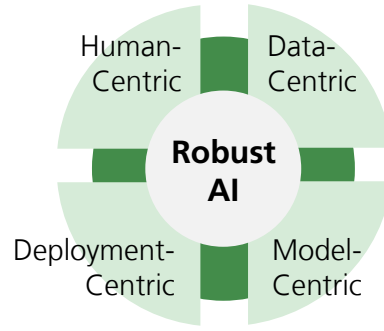


Figure 3.4: Proposed holistic approach to robust AI.

process diverse and complex medical data, deliver accurate predictions, and maintain reliability even in high-stakes environments such as critical care.

On the other hand, robustness at the social level addresses the broader societal implications of AI deployment. It considers the AI system’s interaction with the environment in which it operates, including the ethical, legal, and cultural norms that govern its use. In healthcare, this means that AI systems must be designed with sensitivity to patient rights, equity, and the trust of healthcare providers and patients. Furthermore, caring about social concerns ensures that the AI system’s integration into clinical practice respects patient autonomy, protects privacy, and is transparent in its decision-making processes.

Achieving robust AI in healthcare requires a holistic approach that integrates advancements in data management, model development, human-centric design, and deployment strategies. Each domain contributes to the resilience and reliability of AI systems. In this work, we reconsider how the AI development journey takes part and propose a framework that categorizes the strategies to develop robust AI into four key approaches: human-centric, data-centric, model-centric, and deployment-centric. Figure 3.4 depicts the four approaches in a circular scheme to convey how improving robustness in DSSs is an iterative process.

These constitute the taxonomy we will use as inspiration to answer the research questions identified throughout this thesis.

We frame our contributions for robust clinical DSSs mainly in the first three approaches, as the deployment of the proposed solutions was either left for future work or studied in general terms, specifically in Chapter 4.

Human-Centric Approaches

Human-centric AI in healthcare emphasizes the critical role of clinicians, patients, and other stakeholders in designing, developing, and deploying AI systems. The primary goal is to complement, rather than replace, human decision-making by ensuring that AI tools enhance clinical workflows and align with healthcare providers’ and patients’ needs and expectations [319].

This approach requires a strong focus on patient-centric design [42], where systems

are developed with a deep understanding of clinical contexts and the specific needs of end-users. Ethical considerations like fairness, transparency, and accountability are guided by frameworks like the Ethics Guidelines for Trustworthy AI [130].

To be truly effective, human-centric AI must incorporate human-in-the-loop mechanisms, allowing clinicians to provide continuous feedback. Moreover, explainable AI (XAI) is important in this domain, as it offers transparency in AI decision-making processes, promoting trust and facilitating the acceptance of these technologies among healthcare professionals [103]. Ultimately, human-centric AI aims to enhance clinicians' capabilities, reduce cognitive load, and ensure that the final responsibility for patient care remains with human experts.

Data-Centric Approaches

Data-centric AI focuses on ensuring that the quality, diversity, and integrity of data used to train and deploy AI systems are optimized to support robust performance in healthcare settings. Given that medical data is often heterogeneous, incomplete, and prone to various forms of bias, a data-centric approach involves rigorous data preprocessing, including cleaning, normalization, and handling missing values to ensure integrity and interoperability [357].

The diversity of datasets is also crucial for developing models that generalize effectively across different patient populations and healthcare environments [362]. Strategies such as data augmentation, creating synthetic datasets, and including underrepresented groups in training sets are essential for enhancing a model's robustness and fairness. By building a solid foundation with high-quality data, AI systems can more reliably support clinical decision-making across various scenarios, ensuring that the insights generated are accurate and equitable [173].

Model-Centric Approaches

Model-centric AI has gathered the most research effort and focuses on developing advanced algorithms and architectures designed to handle the inherent complexity and variability of data [205]. In clinical DSSs, this involves creating models capable of processing and integrating multiple data modalities into a cohesive and comprehensive framework [5]. Techniques like data fusion and longitudinal modeling are particularly valuable in this context, as they allow AI systems to capture and analyze the full spectrum of patient information over time [57].

Furthermore, model-centric AI emphasizes the importance of interpretability and transparency, ensuring that models are accurate and understandable to clinicians [286]. This is crucial for gaining the trust and acceptance of AI-based DSSs in clinical practice. Flexible and adaptable architectures, such as transformers and NNs, enable continuous learning and adaptation to new data and domains, ensuring that models remain effective in dynamic clinical environments. Techniques such as fine-tuning and transfer learning

from foundational models [362], particularly those used in NLP or image recognition, can be leveraged to accelerate the development of specialized healthcare models, thereby enhancing their robustness and applicability in real-world settings.

Deployment-Centric Approaches

Deployment-centric AI addresses the ongoing challenges associated with implementing AI systems in real-world healthcare environments [89]. This approach emphasizes the need for AI models to be not only technically robust but also operationally viable within the constraints of clinical settings. Scalability is a crucial consideration, as systems must be capable of handling large volumes of data and processing demands without compromising performance or accuracy [192].

Moreover, deployment-centric AI requires rigorous validation and continuous monitoring to ensure that models perform consistently across diverse healthcare settings and patient populations [264]. This involves establishing mechanisms for real-time feedback and updates, which allow for the continuous improvement of AI systems in response to new data and evolving clinical practices [17].

Security and privacy are also critical components in this domain, as the deployment of AI systems must comply with regulatory standards and ensure the protection of patient data from breaches and misuse [21]. Ultimately, deployment-centric AI aims to bridge the gap between research and clinical practice, ensuring that AI systems are effectively integrated and sustained in clinical DSSs to improve patient outcomes and enhance the overall quality of care.

3.4 Ethical and Legal Frameworks for Responsible and Trustworthy AI

Advancements in AI-based clinical DSSs promise more accurate diagnostics, personalized treatment plans, and efficient healthcare delivery. However, the complexity of real-world medical data, the intricacies of clinical workflows, and the need to integrate DSSs into existing healthcare systems present significant obstacles. Moreover, the successful deployment of these systems depends not only on overcoming technical barriers but also on ensuring that AI is developed and applied in a responsible and trustworthy manner.

Responsible AI encompasses the development and utilization of AI systems that prioritize societal benefits while mitigating risks and adverse effects [316]. Central to this practice are the accountability, responsibility, and transparency (ART) design principles, which ensure that AI systems align with human values and ethical standards [79]. These principles align with responsible innovation strategies used in several areas. They are crucial for fostering trust and ensuring that AI technologies are developed in a manner that is fair and just in consideration of social contexts, addressing concerns about bias and discrimination [49].

Ethical Principles	Key Requirements	
Human Autonomy	Human Agency and Oversight	
Prevention of Harm	Technical Robustness and Safety	
Fairness	Privacy and Data Governance	Transparency
Explicability	Diversity, Non-Discrimination, and Fairness	
	Societal and Environmental Well-Being	Accountability

Figure 3.5: Ethical principles and key requirements for Trustworthy AI, as defined in the EC’s Ethics Guidelines for Trustworthy AI [130].

In the same line of concern and presenting overlapping definitions, Trustworthy AI focuses on ensuring that new technologies are effective and also aligned with societal morals and ethical principles [130]. Trustworthiness in AI extends beyond the system’s technical capabilities to encompass the entire socio-technical context in which the AI operates. This means that AI systems must be lawful, adhering to all relevant regulations, ethical, safeguarding fundamental rights and moral values, and robust, capable of withstanding technical failures and preventing unintended consequences [130]. Trustworthy AI demands a holistic approach, where trust is built not only in the technology itself but also in the processes, actors, and institutions involved in the AI system’s life cycle.

As technologies evolve, robust ethical guidelines and regulatory frameworks become increasingly essential to mitigate the risks associated with their use. The EU has been addressing these issues by creating a framework of tools to guide the development of AI systems that meet the EU values on respect for human dignity, freedom, democracy, equality, the rule of law and rights [144]. This effort is significant given the rapid pace of AI innovation compared to the slower development of societal and regulatory responses.

3.4.1 Ethical Principles and Key Requirements for Trustworthy AI

The European Commission (EC)’s High-level Expert Group on AI proposed in 2019 a Framework for Trustworthy AI, which includes three major components: lawful, ethical, and robust. The ethical foundation for AI is schematized in Figure 3.5 and is anchored in four core principles, as outlined in the Ethics Guidelines for Trustworthy AI [130]. These principles aim to ensure that AI systems are developed and deployed in ways that align with fundamental rights and societal values:

- **Respect for Human Autonomy:** AI systems should enhance human agency and autonomy, not replace or diminish them. In healthcare, this principle emphasizes the need for AI tools to assist clinicians in decision-making without the risk of superseding human decisions.

- **Prevention of Harm:** AI systems must be designed to avoid causing any form of harm, whether physical, psychological, or social. This includes preventing harmful biases, ensuring patient safety, and protecting against misuse of AI technologies in clinical contexts.
- **Fairness:** Fairness requires that AI systems promote equity and avoid bias, ensuring that the benefits of AI are distributed fairly across all societal groups. AI should not perpetuate or amplify existing inequalities in healthcare, such as those based on socioeconomic status, race, or gender.
- **Explicability:** To foster trust in AI, transparency and explainability are essential. Healthcare providers and patients must understand how AI systems reach their decisions. This ensures accountability and enables informed decision-making based on AI recommendations.

These principles are further operationalized through seven key requirements, presented in Figure 3.5, which provide a practical framework for implementing Trustworthy AI throughout its life cycle.

International organizations, such as the Organisation for Economic Co-operation and Development (OECD) [234], International Organization for Standardization (ISO)¹, World Health Organization (WHO) [65], and Institute of Electrical and Electronics Engineers (IEEE)², have also provided complementary standards and frameworks. For instance, the OECD emphasizes inclusive growth and sustainability. In contrast, the ISO focuses on standardization to ensure safety and performance in AI systems for different organizations and applications. These international standards aim to guide the responsible development and deployment of AI technologies worldwide.

Although not legally binding, the USA issued in 2022 the "Blueprint for an AI Bill of Rights" [339]. This framework aims at protecting individuals from potential harms associated with AI systems. The document outlines five fundamental principles designed to ensure that AI technologies are developed and used in ways that align with democratic values and respect individual rights, including the right to be protected from unsafe or ineffective systems, the right to be free from algorithmic discrimination, the right to data privacy, the right to know when an AI system is being used and to understand its implications, and the right to opt-out of AI-based decisions in favor of human alternatives. In line with AI guidelines, this document serves as a guiding document for policymakers, AI developers, and companies to promote the responsible and ethical use of AI.

¹ISO/IEC 23053:2022 (Framework for ML-based Systems) and ISO/IEC TS 4213:2022 (Assessment of ML Classification Performance).

²IEEE 7000-2021 (Model Process for Addressing Ethical Concerns During System Design), IEEE 7001-2021 (Transparency of Autonomous Systems), IEEE 7002-2022 (Data Privacy Process), IEEE 7003 (Draft - Algorithmic Bias Considerations), IEEE 7010-2020 (Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being), IEEE 7012-2021 (Standard for Machine Readable Personal Privacy Terms), IEEE 7014-2021 (Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems) and others in <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/> (accessed on 04/09/2024).

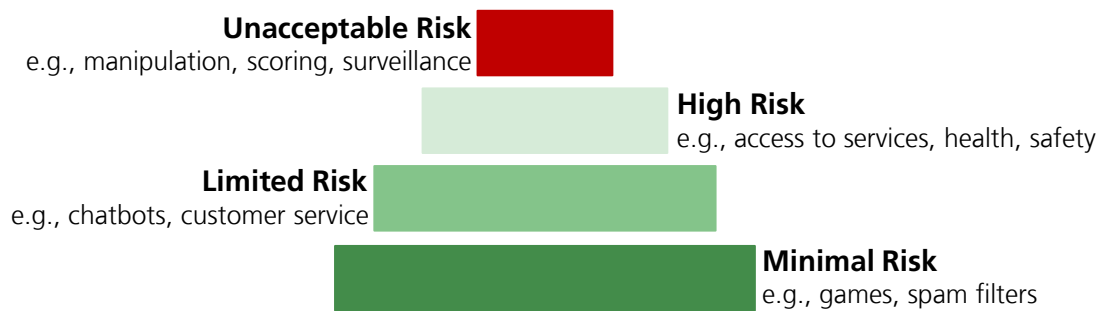


Figure 3.6: Risk levels in the AI Act [92].

3.4.2 Regulatory Frameworks

Although the concept of AI is not new, regulation in the area has only recently been pushed forward since the recent developments. In 2019, the OECD issued a Recommendation on AI [234], which set the first international standard in the form of soft law in an attempt to propose responsible innovation and trust in AI while ensuring respect for human rights and democratic values.

Particularly in healthcare, data protection is crucial. The GDPR [93] of the EU sets the gold standard for how personal data should be collected, stored, processed, and shared. Enforced in 2018, GDPR has profoundly impacted the development and deployment of AI systems in healthcare, where protecting sensitive patient data is critical.

Under GDPR, healthcare providers and AI developers must ensure that data processing activities are transparent, secure, and compliant with the patient's right to privacy. This involves obtaining explicit patient consent to use their data, ensuring data minimization by only processing the data necessary for a specific purpose, and implementing robust security measures to protect against breaches. Additionally, GDPR enforces the "right to be forgotten," allowing individuals to request the deletion of their personal data, which poses a unique challenge for AI systems that rely on large datasets for training and continuous improvement [360]. In healthcare, compliance with GDPR is vital to maintaining patient trust by respecting confidentiality and responsible data usage.

The AI Act

The EU's AI Act [92], proposed in April 2021 and later published in August 2024, represents a significant step toward regulating AI technologies across various sectors, including healthcare. The AI Act introduces a risk-based approach to AI regulation, classifying AI systems into four categories, as shown in Figure 3.6:

- **Unacceptable Risk:** AI systems deemed to pose a significant threat to fundamental rights, safety, or democratic processes are classified as unacceptable and are prohibited. Examples include AI systems that manipulate human behavior or exploit vulnerabilities.

- **High Risk:** AI systems classified as high-risk, such as those used in healthcare, safety, and environment, must meet stringent requirements before deployment. These requirements include rigorous testing, documentation, transparency, and continuous monitoring to ensure safety and effectiveness. For instance, AI systems involved in diagnosing diseases or recommending treatments fall into this category and must comply with specific standards to prevent harm and ensure robust performance.
- **Limited Risk:** AI systems that present risks related to impersonation or deception, such as chatbots or AI-based customer service tools, must adhere to transparency obligations, informing users that they are interacting with an AI system.
- **Minimal Risk:** AI systems that pose minimal risk, such as spam filters, are subject to few regulatory obligations under the AI Act.

For clinical DSSs, particularly those categorized as high-risk, the AI Act mandates comprehensive risk management systems, including pre-market assessments, post-market surveillance, and incident reporting. This ensures that AI systems are continuously evaluated and adjusted based on real-world performance, safeguarding patient safety and enhancing trust in AI-based healthcare solutions.

Additionally, the AI Act recognizes the unique challenges posed by general-purpose models, such as the Generative Pre-trained Transformer (GPT) [235], due to their broad applicability and potential impact. As such, these systems are subject to strict requirements to ensure they operate safely and ethically across different use cases. The requirements include robust risk management, transparency measures, and the obligation to continuously monitor and update the system based on its performance in diverse environments.

The concept of Artificial General Intelligence (AGI) is becoming increasingly relevant with the rise of generative general-purpose models. AGI refers to AI systems that possess the ability to surpass human intelligence in virtually all aspects [212], a notion that is gradually transitioning from science fiction to reality as companies like OpenAI, Google DeepMind, and Anthropic openly pursue this goal [169]. While AGI has not *yet* been achieved, critical concerns about its potential risks are gaining attention. These include the loss of human control over such systems, misalignment with human values, and even existential threats to humanity [45]. As AI capabilities continue to advance, future regulatory frameworks must actively address these challenges and ensure that legal mechanisms evolve to keep pace with AI's potential.

AI-based Medical Devices Regulation

Specifically in healthcare, other regulations already consider using AI-based medical devices. The Software as a Medical Device (SaMD) refers to software intended to be used for medical purposes without being part of a physical medical device. The International Medical Device Regulators Forum (IMDRF) defines SaMD as software that is capable of

performing complex clinical functions, such as diagnosing conditions, managing patient care, or driving clinical decision-making processes [142].

In the EU, SaMD is regulated under the Medical Device Regulation (MDR) [214], which came into force in May 2021. AI-based SaMD must undergo a rigorous clinical evaluation to demonstrate its safety, efficacy, and compliance with essential requirements before it can be marketed in the EU. Similarly, in the USA, the Food and Drug Administration (FDA) oversees the regulation of SaMD, including AI-based applications. The FDA's regulatory framework for AI- and ML-based SaMD [323] focuses on a "total product life-cycle" approach, which emphasizes ongoing oversight from pre-market development to post-market performance monitoring.

These regulatory frameworks are essential for ensuring that AI systems in healthcare are not only innovative but also safe, ethical, and effective in improving patient outcomes. As AI continues to evolve, these regulations must adapt, balancing the rapid pace of technological advancement with the need for rigorous oversight and public trust.

3.5 Final Remarks

The integration of AI into clinical DSSs offers immense potential to enhance healthcare outcomes, particularly in aging populations, by making processes more efficient and addressing the economic sustainability of healthcare systems. With clinicians empowered by AI-driven tools and patients benefiting from more informed, responsible, and trustworthy decision-making, the future of healthcare is promising.

In this chapter, we defined clinical DSSs in healthcare, explored their current state, and highlighted their future potential. Despite the promise, widespread real-world application remains limited due to a lack of robustness. A key challenge lies in the nature of medical data, which presents issues related to quality, complexity, compartmentalization, and the imperative for ethical and responsible usage.

In response to these challenges and recognizing the critical need for more robust medical AI that sets the motto of this thesis, we proposed a framework that categorizes the development of AI systems. This holistic taxonomy considers four critical approaches: human-centric, data-centric, model-centric, and deployment-centric AI. Throughout this thesis, we focus on advancing the first three categories, contributing to creating more robust clinical DSSs.

Beyond technical hurdles, there is an urgent need for regulatory frameworks to guide the development and deployment of AI systems, particularly as advancements in AI continue to accelerate and discussions around AGI gain momentum. The EU has taken a leadership role in defining ethical principles for trustworthy AI, culminating in the AI Act, the first comprehensive AI regulation, which other nations are likely to follow. In line with European values, AI research and development must uphold principles such as respect for human autonomy, prevention of harm, fairness, and explainability. As a risk-based regulatory approach, the AI Act focuses on applications rather than specific

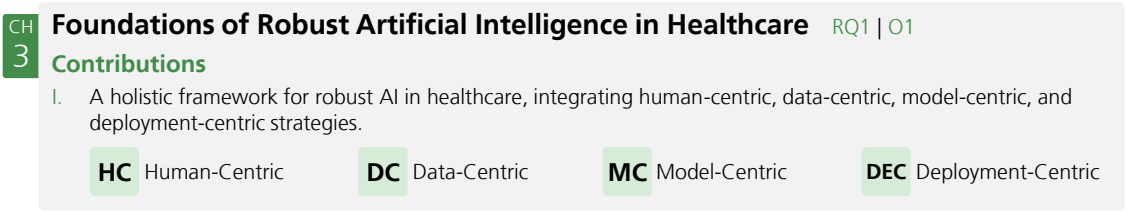


Figure 3.7: Overview of Chapter 3 contributions.

technologies and enforces restrictions on certain high-risk products. In healthcare, regulations governing the certification of AI-based medical devices, along with international standards, are already in place to ensure safety and efficacy.

This chapter has outlined the opportunities and challenges associated with implementing robust AI in clinical DSSs. The main contributions are summarized in Figure 3.7. The following technical chapters (4, 5, and 6) will address these challenges, exploring various approaches to robust AI. Different research directions will be pursued, often combining multiple strategies to improve the robustness of real-world DSSs. We will examine use cases in cardiothoracic surgery, ECG classification, and continuous ICU monitoring, each presenting distinct data types and technical challenges. While risk models for cardiothoracic surgery will primarily involve human-centric and data-centric strategies, with discussions on deployment-centric as well, ECG classification will focus more on model-centric approaches. The complexities of ICU monitoring will require combining data-centric and model-centric strategies to ensure effective and reliable DSSs.

ROBUSTNESS IN CARDIOTHORACIC SURGERY

RISK ESTIMATION: ADDRESSING DATA HETEROGENEITY AND MISSING INPUTS

4.1 Introduction

Real-world data presents significant challenges in developing robust DSSs due to its inherent complexity, including missing values, heterogeneity, and the longitudinal nature of patient information. These challenges can severely impact the accuracy and reliability of predictive models, making it essential to devise strategies that can effectively manage and mitigate these problems. Cardiothoracic surgery is a prime example where these challenges are particularly pronounced, as the variability in patient data and the occurrence of incomplete records demand robust approaches to ensure accurate risk estimation and patient outcomes.

Cardiothoracic surgery involves the heart, lungs, and other thoracic organs. Vervoot et al. [327] estimate an annual prevalence of 123.2 cardiac-related procedures per 100,000 population in high-income countries. Although critical to improving patients' quality of life [68], open-heart procedures present high risks and are prone to severe complications [22]. Despite technological advancements and improvements in clinical standards over the past decades, these procedures remain complex and challenging.

Complications during surgery or hospitalization are common. Cardiovascular, pulmonary, and renal complications remain the major causes of morbidity in cardiac surgery patients, with reports of approximately 4% hospital mortality [22, 228]. Risks extend into the postoperative period. Several studies have reported readmission rates ranging from 13% to 17% in the month following these interventions [165, 276, 290], and up to 30% during the first year [58, 83].

Accurate risk prediction in cardiac procedures has been vital in enhancing patient outcomes and optimizing resource allocation. By identifying high-risk patients before surgery, clinicians can implement targeted interventions to mitigate potential complications. Risk scores have been used in clinical practice, with the EuroSCORE [228] and the

Society of Thoracic Surgeons (STS) [232, 288] risk scores being the gold standards in the field. These models have been widely used to estimate the probability of adverse events based on patient characteristics and clinical factors.

Though valuable, traditional risk prediction models suffer from limited robustness when applied to real-world clinical settings [208]. These models' static nature and reliance on well-curated datasets do not account for the dynamic and heterogeneous nature of real-world medical data. Clinical teams are subject to insufficient human and technical resources and cannot invest in a careful and complete data registry. Factors such as the limited data quality and the variability of clinical practices harm the reliable usage of such important tools.

The challenge of increasing robustness in cardiothoracic surgery risk prediction is determinant to support the continuous improvement of patient outcomes. Robust models must effectively handle the inherent challenges of real-world medical data. Predictive tools must provide consistent and reliable risk assessments across diverse clinical settings and be transparent when applied in real-world scenarios.

This chapter approaches the development of robust predictive models for risk estimation in cardiothoracic surgery, focusing on addressing the technical challenges posed by real-world medical data. These risk prediction models will be used to develop a DSS to support the clinical team at Hospital de Santa Marta - Unidade Local de Saúde São José (HSM) in selecting patients for Remote Patient Monitoring (RPM) following the procedure.

We will discuss data processing techniques to manage data heterogeneity based on informed clinical recommendations and innovative imputation methods to handle missing inputs. Additionally, we will use these methods to explore the integration of longitudinal patient data to enhance the accuracy and reliability of risk predictions over time in the face of missing labels. Throughout the research of these topics, we focus on human-centric, data-centric, and deployment-centric approaches to robust AI towards building more resilient and trustworthy models that can be effectively deployed in clinical settings.

4.2 Case Study: CardioFollow.AI

The deployment of AI models in real-world clinical environments is the ultimate goal of research in DSSs. The COVID-19 pandemic underscored the critical need for robust AI to handle medical data, including multimodal, asynchronous, and incomplete datasets. This need for robustness to ensure that models can be reliably applied and deployed in diverse and dynamic clinical environments inspired the research presented in this thesis.

Our work has been conducted as part of the CardioFollow.AI research project, proposed in response to the challenges posed by the COVID-19 pandemic. As cardiothoracic surgery patients were identified as a high-risk group, many surgeries were delayed or performed under emergency protocols to reduce nosocomial transmission of the virus. CardioFollow.AI studied the effectiveness of alternative follow-up methods to minimize hospital visits while ensuring quality care. The proposed solution included an innovative

RPM service to continuously monitor critical parameters throughout patient recovery, assisted by AI models that provide risk assessments as new data becomes available. CardioFollow.AI aimed to mitigate immediate pandemic-related risks and establish a robust framework for managing future pandemics and routine care.

4.2.1 Remote Patient Monitoring

Post-discharge RPM can be a valuable tool to maximize surgery outcomes, such as reducing mortality, preventing hospital readmissions, and improving long-term quality of life [239, 355]. With such a service, clinical teams can follow patients closely through digitally transmitted data from patients' homes [97].

Acknowledging the potential of telemonitoring, CardioFollow.AI integrated a comprehensive RPM service that included an Internet-of-Things (IoT) kit [199]. The solution was designed by Fraunhofer AICOS (FhAICOS) and included a smartwatch, sphygmomanometer, scale, and a smartphone with the CardioFollow.AI app, enabling continuous data collection. BEATNIK, a web-based platform developed by the Value for Health CoLAB (VOH.CoLAB), was used to allow clinicians to monitor patients remotely, displaying patient information and history of medical parameters, as well as rule-based alerts following anomalous registries [198].

In a clinical study conducted during the project, the RPM service was implemented in HSM, where patients were closely monitored during the three months following hospital discharge. The interval was chosen to balance the efficient monitoring of adverse events with the availability of resources for high-risk patients. Patients performed daily measurements during the first month and weekly registries afterward. Since RPM services require extra resources compared to the standard of care, namely equipment and nursing and technical teams to follow up with patients, AI-assisted patient selection can optimize the allocation of resources.

4.2.2 Data Collection

RPM is one of the sources of clinically valuable data available in the project, focused on the postoperative period. In addition, the Cardiothoracic Surgery Department of HSM has been gathering exceptionally rich medical data throughout the patient journey. In collaboration with researchers from NOVA Medical School (NMS), the clinical and nursing teams implemented a patient follow-up program during the first postoperative year. From 2008 to 2021, 8,933 patients participated in this unique follow-up program in Portugal. This initiative generated a comprehensive database containing patients' clinical data and quality of life metrics, including details on medication and lifestyle modifications gathered through periodic in-person and phone appointments [68]. Figure 4.1 schematizes the data collected throughout the patient journey.

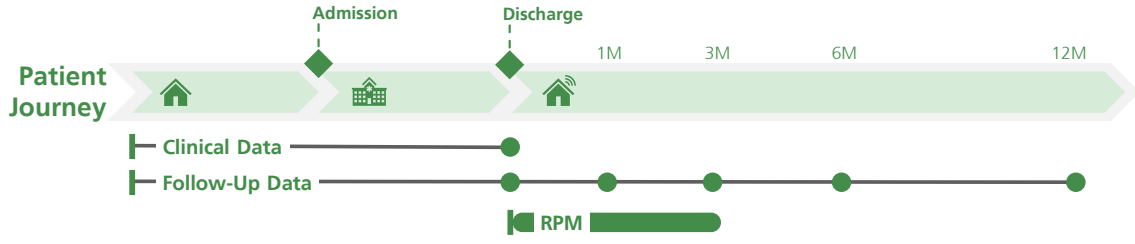


Figure 4.1: Scheme of patient journey. RPM denotes the Remote Patient Monitoring period.

4.2.2.1 Inpatient Clinical Data

The largest data source in HSM includes patient variables related to the pre- to post-surgery periods, up to hospital discharge. The inpatient dataset adheres to the European Association for Cardio-Thoracic Surgery (EACTS) Adult Cardiac Database fields, a widely recognized standard for data collection in cardiothoracic surgery [128]. The data collection form includes 105 variables that capture critical information about patients' demographics, pre-surgery conditions, specifics of the surgical procedures, duration of hospital stay, and in-hospital post-surgery complications. The extensive list of variables is available in Table A.1. Data collection has been conducted by the clinical team since 2008, which culminated in a database including all patients intervened by the Cardiothoracic Surgery Department at HSM.

The inpatient database encompasses many clinical variables that supported this research. It records basic demographic details such as age and gender, along with patients' pre-surgery health status, including comorbidities and clinical history. The database also details the types of surgeries performed and their specific characteristics. Additionally, it tracks the length of stay in different wards during hospitalization and any complications that occurred after surgery. A general summary of the inpatient database is available in Table 4.1.

Table 4.1: Inpatient dataset summary.

	Male	Female	Total
Number of Patients	5,615 (62.9%)	3,316 (37.1%)	8,933
Number of Surgeries	5,903 (62.7%)	3,509 (37.3%)	9,415
Number of Procedures	6,190 (61.1%)	3,943 (38.9%)	10,136
Age (years)	64.9 $\pm$ 12.5	67.1 $\pm$ 14.0	65.7 $\pm$ 13.1

4.2.2.2 Outpatient Follow-Up Data

An additional data collection protocol, coordinated by the nursing team, has been implemented at HSM since 2011. A summary of this dataset is presented in Table 4.2. This process aims at measuring the patients' recovery and quality of life after surgery. A set

of questionnaires is applied to the patients before the procedure, on the third day after hospital discharge, and then at 30, 90, 180, and 365 days after the surgery date. At the end of each period, a nurse contacts the patient by phone to ask the relevant follow-up questions.

The questionnaires include multiple categories of information: physical activity, professional and socioeconomic status, pain assessment, general health status, surgical site interventions, nutrition, therapeutic adherence, vital signs monitoring, family and social support, and potential reference for a hospital appointment. This allows the nursing team to gather data on patients' evolution after surgery.

In addition, the 36-question Short Form Health Survey (SF-36) (second version) [101, 333] is performed before surgery and at 90, 180, and 365 days after to measure health-related quality of life and the long-term impact of medical interventions. It includes 36 questions that assess eight key health domains: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to emotional issues, general mental health, social functioning, vitality, and general health perceptions.

Table 4.2: Outpatient dataset summary.

	Male	Female	Total
Number of Patients	3,349 (63.2%)	1,947 (36.8%)	5,296
Number of Surgeries	3,421 (63.1%)	997 (36.9%)	5,418

Finally, a rich set of complications is annotated at the end of each interval, which can be used as ground truth outcomes for risk prediction models. Death records were verified through the national death certificate system in the case of unsuccessful phone call attempts. Table 4.3 details the considered complications, ranging from light to severe, according to the classification by the clinical team and their prevalence among the registered complications in the dataset. A total of 1,738 complications were identified, among which 834 (48.0%) were considered severe.

4.2.2.3 Remote Patient Monitoring Data

Clinical parameters are continuously recorded during the telemonitoring phase to provide detailed temporal information on each patient's health status. This data includes vital signs such as weight, blood pressure, heart rate, and physical activity, which are measured by the patients daily or weekly with the provided IoT devices that immediately send the parameters to the CardioFollow.AI app and then to a secure cloud infrastructure. Patients take periodic photos of the suture and self-report symptoms through an intelligent natural conversation module, contributing further insights into their daily health and well-being. These data points offer valuable temporal information to help detect complications early or adjust treatment plans promptly.

Table 4.3: Registered complications in HSM and relative frequency among the total number of complications.

Complication	Type	No. Registered	Relative Frequency
Superficial infection (dressing)	Light	563	32.4%
Atelectasis, pneumonia, pleural effusion, or pneumothorax requiring drainage	Light	45	2.6%
Fibrillation or arrhythmia requiring treatment	Light	29	1.7%
Complication not requiring ICU admission	Light	267	15.4%
Stroke with deficit persisting for more than 72 hours (documented by CT/MRI)	Severe	33	1.9%
Sternal infection or instability	Severe	32	1.8%
Angioplasty	Severe	23	1.3%
Endocarditis and/or sepsis	Severe	17	1.0%
Permanent pacemaker	Severe	13	0.7%
Myocardial infarction	Severe	5	0.3%
Acute respiratory distress syndrome	Severe	1	0.1%
Dialysis or hemofiltration	Severe	1	0.1%
Complication requiring surgery (cardiac or not)	Severe	197	11.3%
Complication requiring ICU admission	Severe	87	5.0%
Sudden death	Severe	425	24.5%

Despite the potential of RPM data to provide real-time insights and improve patient outcomes, the primary focus of this research was on leveraging the inpatient and outpatient data modalities, as these offer a richer context and more extensive datasets. The study involved a limited number of participants, indicating potential for future research.

4.2.3 Artificial Intelligence Tasks

The clinical team identified several critical tasks for AI models to enhance patient care in cardiothoracic surgery. These tasks are motivated by the need to improve patient monitoring and risk prediction accuracy and efficiency. The usefulness of AI-based DSSs spans the entire patient journey, which can be divided into three main groups:

- **Pre-surgery:** The period before the procedure comprises two main tasks. In line with gold standard tools, pre-surgery risk prediction can provide insights for clinicians to personalize the surgery plan and recovery stage. Also, optimal resources can be allocated by estimating the required hospitalization days and identifying opportunities for early discharge. Accurate pre-surgical predictions ensure that each patient receives the appropriate level of care, optimizing hospital resources and improving patient outcomes.

- **Post-surgery:** After the surgical procedures, risk prediction models can be used to identify patients that would benefit the most from RPM. By analyzing post-surgical risk factors, the system can identify high-risk patients and enroll them in RPM programs, ensuring that those who need close monitoring receive it. This selective enrollment is crucial for maximizing the benefits of telemonitoring and preventing complications.
- **Post-discharge:** A successful procedure does not guarantee a low risk of complications. Therefore, a risk monitoring system can enable prompt medical intervention by continuously analyzing patient data and identifying early signs of complications. This assistance impacts long-term patient health outcomes and reduces the burden on healthcare facilities.

These tasks inspired the work presented in this chapter. While Sections 4.3 and 4.4 address the second task, Section 4.5 addresses the third. We collaborated with clinicians during data preparation and objectives definition to develop a DSS trusted by doctors for their transparency, performance, and reliability.

4.3 Developing Risk Models for Cardiothoracic Surgery¹

4.3.1 Objectives and Significance

Risk prediction models have been used in cardiothoracic surgery for many years [241]. The EuroSCORE was initially introduced in 1999 and later replaced in 2012 by the EuroSCORE II. These scores were designed to predict in-hospital mortality and apply to various procedures [6]. In contrast, there are several STS Scores specifically tailored to particular cardiac procedures that aim to predict not only mortality but also various postoperative complications such as kidney failure, stroke, and deep wound infection, among others [6, 232, 288]. These scores assess outcomes up to hospital discharge or within a 30-day postoperative period.

However, the risk of complications extends into the first few months after surgery, with up to 30% of patients requiring readmission within the first year [83]. Caruso and Zadra [58] demonstrate with statistical significance that several complications tend to occur beyond the first month but mainly before the third. Furthermore, the authors of the EuroSCORE II highlight an additional 0.9% mortality risk when extending the observation period to 90 days, compared to the initial 30 days, which already shows a 0.6% increase compared to the in-hospital mortality risk [228]. A significant limitation identified in the study is that most participating centers do not track outcomes beyond hospitalization. This

¹This Section is based on the publication “Predicting Post-Discharge Complications in Cardiothoracic Surgery: A Clinical Decision Support System to Optimize Remote Patient Monitoring Resources” by **Ricardo Santos**, Bruno Ribeiro, Inês Sousa, Jorge Santos, Federico Guede-Fernández, Pedro Dias, André V. Carreiro, Hugo Gamboa, Pedro Coelho, José Fragata, and Ana Londral, published in the *International Journal of Medical Informatics* (2024) [280].

reduces the model's applicability, hence the short 30-day outcome window of EuroSCORE II. Given the restricted predictive intervals and the often overlooked complications beyond mortality, considerable scope exists for improvement.

Asserted the potential of RPM in cardiothoracic surgery and the limitation of resources that such service still poses to clinical care, **this work proposes to design a clinical DSS to select patients for RPM considering limited resources.**

Since standard risk scores have limitations and are not specific for this purpose, a ML model was developed to predict the risk of complications within 90 days after discharge from cardiac surgery. The construction of a robust ML pipeline carefully addressed challenges typical in medical AI, such as data quality, extreme imbalance, and trustworthiness [53]. The proposed DSS was applied in HSM to test the validity of RPM in cardiac surgery in the scope of project CardioFollow.AI.

The hypothesis under validation is that the optimal value of such a digital service is achieved when applied to patients at higher risk of complications. As detailed in Section 4.2.1, the RPM service is designed to last for three months following patient discharge; therefore, the risk prediction system needs to predict complications within this interval.

4.3.2 Literature Review

4.3.2.1 Traditional Risk Models

Risk models in cardiothoracic surgery, like those in other medical specialties, are based on more straightforward statistical and regression methods. They use variables about the surgery and the patient, including demographic data, comorbidities, and medications. Such models primarily aim to estimate the risk of the surgery during its planning.

The EuroSCORE is one of the most used scores in clinical practice [229]. Launched in 1999, the risk score was designed to stratify early mortality risk in cardiac surgery patients. Developed using approximately 1,500 European patients, the score is based on several expert-selected risk factors that fall into three categories: patient-related, cardiac, and operation-related factors. These factors were chosen for their objectivity, credibility, availability, and resistance to falsification. Logistic Regression (LR) was used to adjust their weights.

In 2012, the same research group published an updated version of the original risk model, the EuroSCORE II [228]. This update used a more recent dataset of 22,381 patients from 154 hospitals across 43 countries, collected over three months in 2010. The authors noted that improvements in care standards had caused the original model to overestimate early mortality risk. While the original risk factors were still considered, new important variables were also included.

Clinicians may also use models from the STS, which include many patient-related variables [247, 288]. The STS Adult Cardiac Surgery Database [100], established in 1989, contains records from over 1,000 hospitals in the USA and aims to improve patient safety and monitor the quality of cardiothoracic surgery. The initial risk models based on this

data were developed in 2003, using records from 1997 to 1999 [292]. These models have been updated twice, in 2008 and 2017, to reflect advances in medical practice and healthcare conditions. They predict severe outcomes following Coronary Artery Bypass Graft (CABG), valve procedures, or both. Each model targets different outcomes, such as operative death, permanent stroke, renal dysfunction, dialysis, reoperation, prolonged ventilation, and deep sternal wound infection. These multivariate LR models utilize various variables, including patient demographics, past conditions, surgery details, and specific lab results.

In a study conducted by Ad et al. [6] using data from a single center in the USA from 2001 to 2016, the predicted operative mortality was 2.7%, 3.3%, and 7.8% for the STS risk score, EuroSCORE II and EuroSCORE I, while the observed was only 1.8%. This translates into AUROCs of 84.6%, 84.4%, and 81.9%, in the same order.

Although commonly used, the EuroSCORE and STS models rely on assumptions about data characteristics, such as the independence of training cases and the absence of multicollinearity. These assumptions, if unmet, can compromise model performance [152]. While these models are effective for short-term mortality risk assessment and quality of care indicators, broader evaluations of patient status before, during, and after surgery require new approaches.

4.3.2.2 Machine Learning-based Risk Models

Over the last decade, researchers have focused on more sophisticated ML-based solutions to approach specific risk-prediction tasks associated with cardiothoracic surgery.

Comparing the performance of the EuroSCORE with various ML algorithms, Allyn et al. [14] experimented with SVM, Extreme Gradient Boosting Machine (XGBoost), RF, NB, and an ensemble model. The ensemble model achieved the best results in predicting in-hospital mortality risk following surgery, outperforming both the other ML methods and the EuroSCORE with an AUROC of 79.5%. In this study and specific data cohort, the EuroSCOREs I and II attained AUROCs of 71.9% and 73.7%, respectively.

In a similar study, Mortazavi et al. [223] applied different ML algorithms to predict general adverse events, respiratory failure, and infection following cardiac surgery. They achieved AUROC scores of up to 83% in assessing individual complications.

Analogously, Fan et al. [96] combined preoperative and intraoperative factors in multiple ML models to predict 30-day post-surgery mortality. Compared to traditional risk scores, advanced strategies outperformed clinical scoring systems with personalized risk profiles.

By incorporating intraoperative factors, Fernandes et al. [99] aimed to predict operative mortality, including deaths during hospitalization or within 30 days post-discharge following cardiothoracic surgery. They focused on patients undergoing Cardiopulmonary Bypass (CPB) and tested various approaches: preoperative and intraoperative factors outside CPB, factors during CPB, and factors throughout the entire surgery. Using different

ML methods, they obtained AUROC scores up to 87%. However, the models showed a high false alarm rate with low precision, close to 10%, a common consequence of high class imbalance.

A study by Kilic et al. [167] predicted the risk of operative mortality after cardiac surgery using an XGBoost model, comparing its results with the STS risk score. Trained on over 1,100 patients from a single hospital using 10-fold CV, both models used the same variables for a fair comparison. Serum creatinine, weight, and age were the most critical risk factors. The XGBoost model outperformed the traditional STS risk model, with an F1-Score of 28.1% compared to 23.0%. Despite similar overall metrics, the low correlation between the models indicated different prediction error patterns.

Benedetto et al. [31] evaluated various ML methods for predicting in-hospital mortality using data from 28,761 cardiac procedures in the United Kingdom (UK)'s National Adult Cardiac Surgery Audit dataset. They used EuroSCORE variables and found that LR outperformed more advanced techniques like NNs. Despite overlapping confidence intervals between different strategies, simpler and more transparent models like the LR may still be preferable due to their comparable performance.

Efforts to optimize traditional risk models using similar data have often faced challenges like highly imbalanced datasets and low precision, leading to an overestimation of risk. However, these models can still benefit clinical settings when considering the costs of lives and money. There remains significant room for improvement in robustness across all steps of the ML pipeline, from feature engineering to hyperparameter selection, which this work addresses.

4.3.3 Methodology

The design of this risk prediction system to select patients for RPM was a collaborative effort conducted in close partnership with surgeons and nurses. Regular meetings and discussions enabled a continuous exchange of insights and validation of assumptions, ensuring that the system aligns with medical expertise and effectively addresses real-world clinical challenges.

4.3.3.1 Dataset

This study utilizes the retrospective inpatient database presented in Section 4.2.2.1, duly anonymized. Data collection adhered to the Declaration of Helsinki, and the study protocol received approval from the hospital's ethical committee (INV 303). No informed consent was required due to the retrospective and observational nature of the study.

Complications were used as labels in developing the risk model, retrieved from the dataset described in Section 4.2.2.2. All severe complications were aggregated into a binary classification task. For patients experiencing multiple complications, the most severe was considered. While in-hospital post-surgery events and complications were used as features, the model's goal was to predict post-discharge complications.

The study cohort included all adult patients who underwent open-heart surgery at HSM from 2008 to 2019. However, only surgeries from 2011 onward were relevant since the registry of complications only started that year ($n=5,568$). Exclusion criteria were patients who died during the hospital stay ($n=182$) and those who did not respond to follow-up phone calls ($n=341$).

Each database entry contains a single medical procedure, representing one ML sample. A patient may undergo multiple surgeries, and each surgery may include several procedures (e.g., CABG, valve replacement). This resulted in 5,045 patients, totaling 5,247 surgeries and 5,625 medical procedures. Sample weights were normalized by the number of medical procedures in each surgery. Out of 5,625 samples, 410 were associated with severe complications within 90 days post-discharge, resulting in 7.3% positive samples, indicating a strongly imbalanced dataset. Table 4.4 summarizes the most relevant variables and their statistical comparisons between classes.

4.3.3.2 Data Preparation

Several steps were taken during the preprocessing stage to ensure data quality and relevance. This stage was time-consuming but crucial in the project's scope. We followed an iterative process of proposing new operations, followed by clinical validation. Such a clinician-in-the-loop approach to promoting high-quality standards of data treatment allowed the proposal of a robust risk estimation system.

A two-fold testing strategy was implemented. First, all data from 2019 were allocated to test set A to evaluate the model's generalization ability over time. From the remaining samples (2008-2018), 30% were designated as test set B, while 70% were used for training the ML models. Each patient was assigned to only one set to avoid data leakage. Table 4.5 provides an overview of the sample distribution across these sets. These splits were conducted before any data processing.

Variables with a high rate of missing values often contain implicit knowledge about clinical workflows. For instance, if a procedure is not performed, its associated metrics are not recorded. The clinical team clarified the meaning of missing values for each variable, and corrections were made as necessary. A missing value for many variables, like "Coronary Implant Type," indicates that no implant was used during surgery. These were, therefore, replaced with a new value representing a null outcome, such as "no implant." Conversely, some variables require a specific condition for accurate imputation. For example, the "Most Recent Myocardial Infarction" variable, which categorizes time intervals, only applies if the patient has had at least one myocardial infarction. Otherwise, missing values are replaced with "never". Variables dependent on other conditions retained some missing values, whereas independent variables had all missing data fully imputed. While it is impossible to completely rule out other causes for some missing values, these instances are expected to be negligible among the total samples.

Table 4.4: Dataset summary for each population group. Numerical variables are presented as the mean and standard deviation ($\mu \pm \sigma$), while categorical variables display the prevalence of the most common value (%). * Canadian Cardiovascular Society grading of *angina pectoris*.

	Overall	No Complications	Complications
Population Size (samples)	5,625	5,215	410
Sex (male)	60.8	61.1	56.6
Age (years)	65.7 $\pm$ 13.1	65.5 $\pm$ 13.2	67.0 $\pm$ 11.6
Height (cm)	164.6 $\pm$ 9.3	164.7 $\pm$ 9.3	164.2 $\pm$ 9.5
Weight (kg)	73.8 $\pm$ 14.2	73.8 $\pm$ 14.0	74.4 $\pm$ 16.0
Body mass index	27.1 $\pm$ 4.7	27.0 $\pm$ 4.6	27.7 $\pm$ 5.8
Smoking history (never)	67.2	67.3	65.6
Angina (Grade I*)	48.2	47.8	53.2
Cerebrovascular disease (no)	92.8	93.1	89.2
Chronic lung disease (no)	90.6	91.0	86.9
Diabetes treatment (no)	69.2	69.9	60.7
Hypercholesterolaemia (yes)	66.7	66.9	63.7
Neurological dysfunction (no)	98.0	98.2	95.3
Congestive heart failure (no)	61.1	61.8	52.0
Renal disease (no)	95.0	95.5	88.9
Operative urgency (elective)	95.9	96.2	92.4
Duration of surgery (min),	199.7 $\pm$ 70.1	198.0 $\pm$ 68.0	220.4 $\pm$ 90.3
Procedure groups (valve only)	44.7	44.8	44.3
Last preop. creatinine (μ mol/l)	94.0 $\pm$ 54.6	92.9 $\pm$ 52.6	108.3 $\pm$ 74.7
Ventilated (no)	99.3	99.4	98.0
Bleeding complications (>10 ml/kg at 48h and/or transfusion)	63.2	62.2	75.4
Cardiac complications (no)	70.6	71.9	53.7
CNS complications (no)	94.2	94.6	88.6
Infective complications (no)	98.5	98.8	95.4
Renal complications (no)	96.8	97.5	88.8
Respiratory complications (no)	92.2	93.4	76.8
Preoperative stay (days)	1.5 $\pm$ 6.5	1.5 $\pm$ 6.6	2.1 $\pm$ 4.3
UCI + Intermediate stay (days)	2.6 $\pm$ 3.2	2.4 $\pm$ 2.4	5.1 $\pm$ 8.0
Postoperative stay (days)	8.5 $\pm$ 11.1	8.1 $\pm$ 9.4	13.6 $\pm$ 23.2

Table 4.5: Summary of the train, test A, test B, and complete sets.

	Train Set	Test Set A (2019)	Test Set B (2008-18)	Total
No. Patients	3,104	610	1,331	5,045
No. Surgeries	3,231	625	1,391	5,247
No. Samples	3,479	667	1,479	5,625
% Samples	61.8	11.9	26.3	100
No. Positives	262	49	99	410
% Positives	7.5	7.3	6.7	7.3

The encoding process also considered the clinical significance of the variables. Categorical variables were one-hot encoded, while those with dependent values (e.g., "Diabetes treatment" ranging from none to insulin injection) were subject to ordinal encoding. Dates of previous procedures were binarized to indicate whether they had been performed, thus overcoming high cardinality.

Two additional features were computed for postoperative complications beyond the individual complication types listed in Table 4.4. The first was a binary feature indicating the occurrence of any in-hospital complication, and the second was a count of all registered complication events. These features enhanced the model's understanding of the interaction and co-occurrence of complications.

Variables with more than 90% missing values or the same value in over 85% of entries were removed, as well as those with a Pearson's correlation coefficient above 90% with any other variable. These thresholds were iteratively tested using a baseline RF model. Finally, min-max normalization was applied, and the remaining missing values were imputed using a multivariate iterative imputation model [238]. After this data preparation, 58 variables were used for model training.

4.3.3.3 Training Pipeline

The proposed training pipeline, depicted in Figure 4.2, follows a structured approach to model development and evaluation. The process begins with a group-stratified 5-fold split of the training data for CV. This split allows for the performance comparison of models and their hyperparameters, identification of the most critical features, optimization of the class separation threshold, and selection of the best sample fusion method.

Five distinct models (SVM, NB, DT, LightGBM, and RF) had their hyperparameters tuned using grid search. The AUROC was used as the optimization metric. The hyperparameter ranges are detailed in Table A.2. We adopted a weighting strategy to address the high class imbalance, considering the inverse of the class prevalence [168].

Feature importance was determined through forward feature selection [248] using three ranking methods: (i) statistical analysis of variance, (ii) permutation, and (iii) embedded methods based on impurity for tree-based models. An early stopping mechanism halted

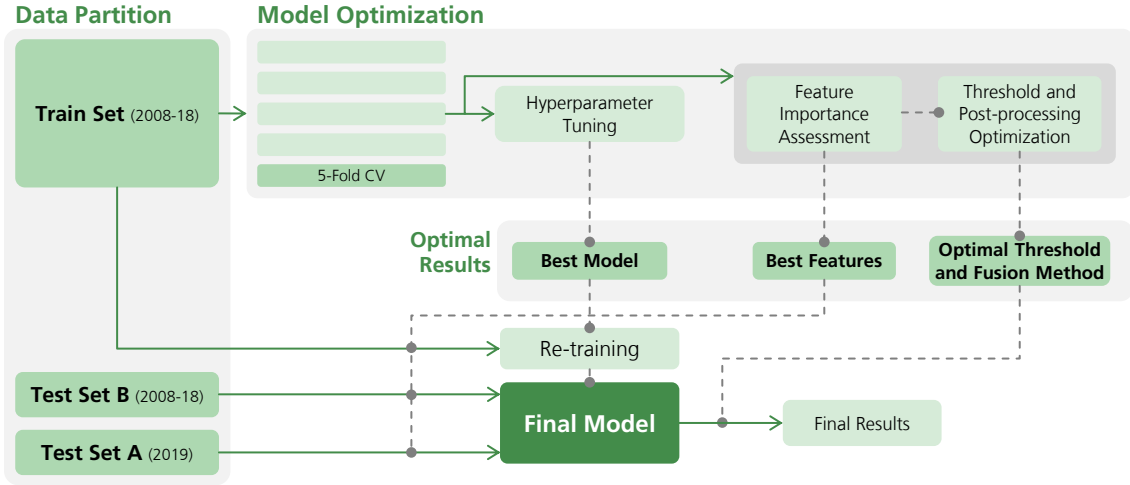


Figure 4.2: Training pipeline proposed for cardiothoracic surgery risk prediction. The green arrows represent the data flow, while the grey dashed lines relate to the passage of parameters between stages.

the iterative cycle after five consecutive iterations with no performance improvements, resulting in the selection of the top k features.

The class separation threshold was optimized in each validation fold based on the weighted average of sensitivity and specificity, with a 2:1 ratio. Clinician feedback informed this decision, emphasizing the need to prioritize sensitivity to accurately identify high-risk patients for RPM. The final optimal threshold was the average of all five folds.

Two post-prediction sample fusion methods (average and maximum) were tested to determine the best approach for handling multiple procedures in the same surgery, which results in multiple predictions but only one expected outcome. After identifying the best hyperparameters, features, class separation threshold, and sample fusion method, each model was re-trained on all training samples using the top k features.

4.3.4 Experimental Results

Deploying ML models in real-world contexts requires rigorous validation, especially when decisions about clinical interventions in human lives are involved. Multiple steps must be considered, ranging from initial tests to subsequent evaluations of clinical outcomes. The proposed model estimates the risk of complications to determine the allocation of resources for RPM, which does not place patients at risk. Instead, it serves as an additional layer of care rather than replacing existing procedures. Nonetheless, we implemented a validation process to assess the pipeline's performance against common challenges, such as its ability to generalize in the face of natural shifts in the population.

The initial step of the model development consists of selecting the best performances during hyperparameter tuning. The mean and standard deviation of the AUROC in the validation sets during CV were used to select the optimal pipelines. Table 4.6 summarizes the optimal models identified through our training pipeline. The RF model, trained on the

top seven features selected via permutation-based ranking, achieved the highest AUROC score. It is noteworthy that the inclusion of additional features did not consistently enhance model performance. 12 of the 13 tested combinations reached their optimal performance with ten or fewer features. Remarkably, a DT model using only four features yielded an AUROC score just 2.9% lower than the best-performing model.

Table 4.6: Pipeline optimization top performances achieved through each model and feature importance method in terms of AUROC, Sensitivity, and Specificity. The mean and standard deviation of all train set's CV folds are presented. The highest value for each metric and parameter is highlighted. The results use the optimal decision threshold and fusion method.

	Importance	k	AUROC (%)	Sensitivity (%)	Specificity (%)
NB	Univariate	9	67.1 ± 2.9	64.0 ± 3.8	61.3 ± 6.8
	Permutation	21	68.6 ± 2.6	62.6 ± 4.5	68.1 ± 6.0
SVM	Univariate	4	67.6 ± 2.6	66.1 ± 10.6	65.6 ± 11.5
	Permutation	5	65.2 ± 1.8	52.9 ± 6.7	68.2 ± 9.2
DT	Univariate	10	66.9 ± 3.4	41.3 ± 22.1	79.5 ± 10.7
	Permutation	4	68.8 ± 2.5	37.9 ± 7.6	85.5 ± 6.7
	Embedded	7	68.2 ± 1.2	47.0 ± 9.3	77.9 ± 6.3
RF	Univariate	9	70.8 ± 2.7	66.2 ± 8.9	68.0 ± 9.5
	Permutation	7	71.7 ± 1.9	67.8 ± 9.3	66.4 ± 8.3
	Embedded	5	71.6 ± 1.9	72.4 ± 7.1	62.4 ± 7.0
LGBM	Univariate	10	70.0 ± 2.6	65.7 ± 4.7	66.8 ± 6.5
	Permutation	8	70.5 ± 2.3	66.3 ± 5.4	66.9 ± 6.6
	Embedded	4	70.4 ± 2.1	65.4 ± 4.3	65.6 ± 6.2

To showcase the limited impact of the number of features in performance increments, Figure 4.3 illustrates the learning curve of the best-performing RF model during the iterative feature selection process. Although all 58 features were evaluated, the model's performance was primarily influenced by the first four features. The early stopping mechanism determined that the optimal number of features was $k = 7$, resulting in an AUROC of $71.7 \pm 1.9\%$.

Following the selection of the best pipeline that optimized the number of features through permutation ranking, Figure 4.4 displays the most important attributes. The ranking highlights a dominance of postoperative information, with only one preoperative feature ("Preoperative Stay") included. The best predictive performance is achieved using three main types of variables: lengths of stay, in-hospital complications, and surgery details, specifically its duration. Two engineered features, i.e., the number of in-hospital postoperative complications and the combined duration of intensive care unit and intermediate stays, appear to enhance the model's performance.

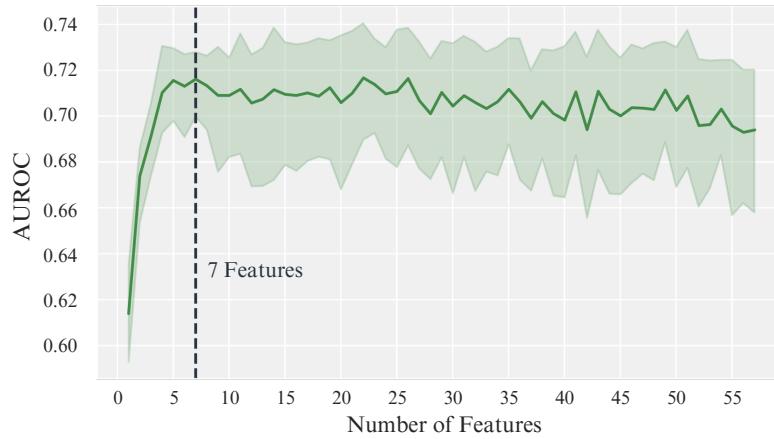


Figure 4.3: Iterative feature selection performance evolution with new features, as given by the AUROC (mean and standard deviation across five folds). The learning curve corresponds to the best RF, trained upon the iterative addition of the next best feature based on a permutation importance ranking.

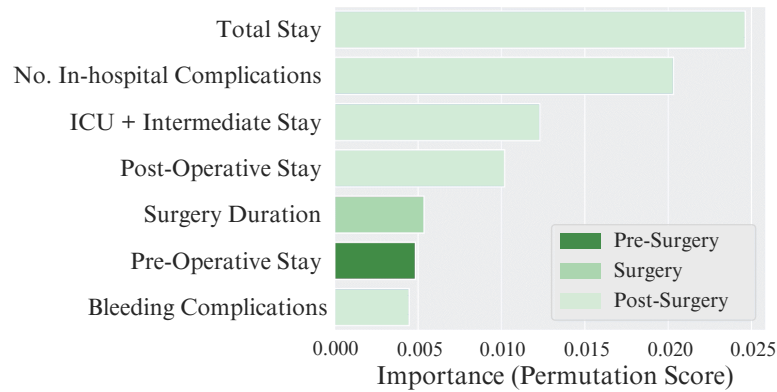


Figure 4.4: Feature set associated with the best-performing model during the optimization phase. Features are ordered according to their importance given by the permutation score. The color scale reflects the nature of the variables.

The optimization phase identified the most effective pipeline based solely on the training set. To estimate the robustness of each model, their generalization capacity was tested using two held-out datasets: test set A (2019) and test set B (2008-2018), as detailed in Section 4.3.3.2. The performance results for each model are presented in Table 4.7.

Overall, tree-based algorithms demonstrated superior performance, achieving AUROC values above 62% on both test sets. In contrast, the SVM and NB models showed a significant drop in performance on test set A, indicating they had overfitted to the specific traits of the 2008-2018 data. Conversely, the RF model maintained AUROC values above 65% on both test sets, demonstrating consistent performance across different temporal windows. Consequently, the RF model was chosen for integration into the DSS developed.

Table 4.7: Final results validated against the test sets, in terms of AUROC, Sensitivity, and Specificity. The highest value for each metric and parameter is highlighted.

	Dataset	AUROC (%)	Sensitivity (%)	Specificity (%)
NB	Train (5-Fold CV)	68.6 $\pm$ 2.6	62.6 $\pm$ 4.5	68.1 $\pm$ 6.0
	Test A (2019)	56.6	41.3	76.5
	Test B (2008-2018)	73.2	60.6	75.8
SVM	Train (5-Fold CV)	67.6 $\pm$ 2.6	66.1 $\pm$ 10.6	65.6 $\pm$ 11.5
	Test A (2019)	56.2	30.4	76.7
	Test B (2008-2018)	69.3	56.4	71.7
DT	Train (5-Fold CV)	68.8 $\pm$ 2.5	37.9 $\pm$ 7.6	85.5 $\pm$ 6.7
	Test A (2019)	62.4	19.6	90.9
	Test B (2008-2018)	68.4	33.0	88.4
RF	Train (5-Fold CV)	71.7 $\pm$ 1.9	67.8 $\pm$ 9.3	66.4 $\pm$ 8.3
	Test A (2019)	65.3	52.2	70.0
	Test B (2008-2018)	69.5	66.0	65.2
LGBM	Train (5-Fold CV)	70.5 $\pm$ 2.3	66.3 $\pm$ 5.4	66.9 $\pm$ 6.6
	Test A (2019)	63.0	43.5	72.4
	Test B (2008-2018)	68.1	64.9	67.8

4.3.4.1 Real Scenario Simulation

In addition to standard robustness evaluation strategies, we implemented a simulation scenario to mimic the real use case of the developed DSS. Test set A (2019) was used to design a weekly RPM resource distribution scenario, evaluating the model’s performance in terms of hypothetical complications in remotely monitored patients. As high-risk patients are prioritized for receiving RPM, a robust patient selection system should ensure that most severe complications occur within the monitored group.

Patients were grouped by the week of surgery, and their risk scores were calculated based on a fixed number of available RPM kits per week. The C-index obtained was compared to that of a random model. Harrell’s C-index, a commonly used metric for risk score evaluation, compares each subject’s predicted score to every other subject with a different outcome, ensuring the model’s risk hierarchy is correct [126]. For binary classification problems, the C-index is equivalent to the AUROC.

The real-world effectiveness of the model was assessed using all surgeries from test set A. In 2019, an average of 11.8 ± 3.7 patients underwent surgery each week. We evaluated how many high-risk patients would have been monitored by distributing RPM kits based on sorted risk scores. The model’s performance was compared to a random distribution, with results summarized in Table 4.8. The proposed risk model achieves better distribution than the random alternative. Also, three kits would be necessary to prevent patients who came to suffer complications from being left out due to a lack of resources.

Table 4.8: Results of weekly RPM kit distribution simulation for all 46 patients who suffered severe complications in 2019. C-index values are the mean and standard deviation of all weeks' risk scoring. The random model followed a uniform distribution, and the results are presented as the mean and standard deviation of 10 runs, thus presenting different C-index when different numbers of RPM kits are considered.

	Model	Patients Monitored	Patients Out due to Model	Patients Out due to Resources	C-Index (%)
2 Kits	Proposed	19	27	2	67.8 ± 25.6
	Random	7 ± 2	37 ± 2	2	48.8 ± 26.6
3 Kits	Proposed	21	25	0	67.8 ± 25.6
	Random	11 ± 2	35 ± 2	0	50.3 ± 27.1
5 Kits	Proposed	28	18	0	67.8 ± 25.6
	Random	18 ± 3	28 ± 3	0	47.8 ± 26.6

4.3.5 Discussion

In the scope of project CardioFollow.AI, the objective of Section 4.3 was to develop a DSS using ML to predict severe complications within 90 days post-discharge for cardiothoracic surgery patients. The RF model emerged as the best performer, achieving AUROCs of 65.3% and 69.5% on test sets A (2019) and B (2008-2018), respectively. Despite the large number of clinical variables available, only four to seven features impacted the final model's performance, most of which were postoperative data. The DSS was built to meet the real-world needs of HSM, proving beneficial in a simulated RPM resource allocation experiment compared to random resource distribution.

4.3.5.1 Long-Term Risk Prediction

The model's ability to predict severe complications or death within 90 days post-discharge demonstrates its robustness, achieving over 65% AUROC in both test sets. The model effectively distinguishes between high-risk and low-risk patients by optimizing the threshold to prioritize sensitivity over specificity following a clinician-in-the-loop approach. Given the limited availability of RPM equipment, it is crucial to consider the probability of risk for each patient compared to the population. Therefore, the AUROC remains the key metric for evaluation.

Going beyond the literature by extending the prediction period to 90 days introduces challenges compared to clinical guidelines and benchmarks, as most cardiothoracic surgery risk models focus on outcomes within 30 days post-intervention. Since patients mostly recover outside the hospital, this more extended period is more likely influenced by external factors. In addition, the population cohort includes both urgent and elective cases and covers various cardiothoracic procedures. While standard risk scores target mortality prediction, the proposed model uses a broader definition of severe events, incorporating

different complications. These complexities justify the lower performance compared to shorter-term studies in Section 4.3.2. However, our results are consistent with [223], who reported a 66% AUROC for general postoperative complications but over a shorter period.

Short-term risk estimation techniques, such as the EuroSCORE [228] and STS scores [288], are widely used in clinical settings for the assessment of immediate surgery outcomes. However, they often overlook long-term outcomes, which are equally important. Studies have shown that adverse outcomes from cardiac surgery can manifest beyond the typical 30-day window, with several authors advocating for a 90-day post-discharge period as a better indicator of hospital performance [131, 153]. A 2016 study found that 40% of readmissions occurred between 30 and 90 days after discharge [104]. This evidence underscores the importance of post-discharge risk prediction, particularly for RPM resource allocation, to mitigate or prevent complications.

4.3.5.2 Post-Discharge Risk Factors

Pre-surgery attributes commonly cited in the literature as relevant for cardiothoracic surgery risk stratification include patient demographics, status, clinical history, procedure urgency, and comorbidities [128]. However, the feature importance assessment revealed that none of these pre-surgery attributes were selected for the final model. At the same time, information collected during and after surgery dominated the importance ranking (see Figure 4.4).

The length of hospital stay is a proxy for the patient's health evolution, influenced by preoperative conditions and unexpected events during or after surgery. While pre-surgery risk factors indicate a potentially adverse response to the intervention, in-hospital complications confirm these scenarios. Additionally, the duration of surgery captures procedural specifics and any intraoperative setbacks.

Postoperative variables effectively incorporate preoperative attributes, rendering the latter redundant while adding crucial context related to short-term incidents. However, annotating postoperative variables requires extra effort, and such models are limited to post-surgery applications. Therefore, it is essential to consider the use case scenarios when evaluating the types of information to be studied.

4.3.5.3 Prediction Interpretability

The robustness of a ML-based DSS is also affected by factors such as the interpretation of outcomes by the end-users, i.e., the clinicians in this use case. To enhance clinical trust, SHapley Additive exPlanations (SHAP) values were used to explain individual predictions [202]. A plot showcasing the importance of each feature, depicted in Figure 4.5, is presented to the clinicians. Instead of using traditional SHAP value plots, each feature's Shapley value was normalized to the sum of the absolute values of all contributions, providing a more intuitive representation. Additionally, the contributions were organized

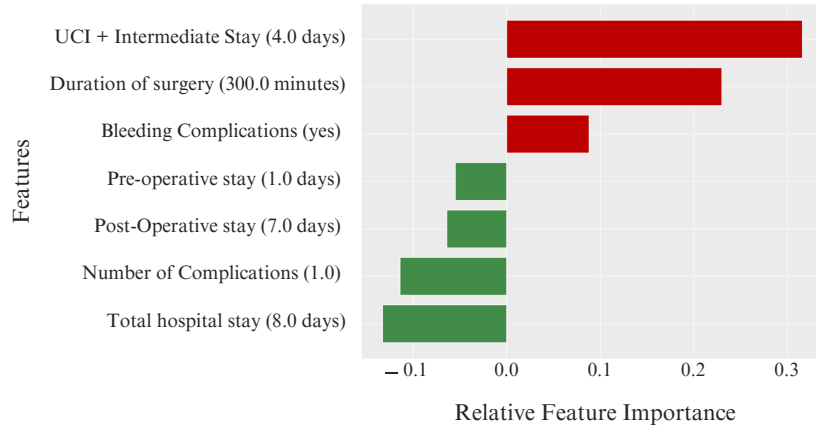


Figure 4.5: Example of Shapley values obtained for each feature. Red bars increase the probability, and green bars decrease it.

based on their impact, distinguishing features that increase risk probability from those that decrease it.

In the example shown in Figure 4.5, the extended ICU stay of four days, which is longer than the overall average (refer to Table 4.4), was the most significant negative factor contributing to the high-risk prediction. The prolonged duration of the surgery, coupled with identified in-hospital bleeding complications, further reinforced this negative prediction. Ultimately, the prediction was classified as high-risk, as the mitigating factors (green bars) showed a smaller cumulative feature importance, which was insufficient to counterbalance the negative influences.

4.3.5.4 Clinical Impact

The hypothetical real-world impact of risk estimation was evaluated in the context of RPM resource allocation. The proposed model consistently outperformed random resource allocation, demonstrating significant potential benefits.

The tested random allocation is only partially realistic. In a setting without a resource allocation system, clinicians would attempt to identify higher-risk patients. This subjective evaluation, the Eyeball Test, can introduce bias and lead to incorrect patient selection, thus not fully leveraging RPM's potential gains [145]. Additionally, clinician estimates could be contextually biased since the dataset was collected from a single hospital.

The effectiveness of the DSS relies heavily on the active involvement of clinicians throughout the design and validation phases. Cardiothoracic surgeons and nurses were integral to the development process, offering valuable insights into data processing, objective definition, label organization, and results validation. This collaborative approach significantly enhances the model's reliability and applicability to real-world clinical scenarios.

Due to monetary and human resource constraints, RPM cannot be universally applied, highlighting the importance of the developed DSS for optimal patient stratification and

equipment allocation. This ensures a scientific, impartial, and balanced method of patient selection. Although this model may not be directly transferable to other populations, the study identified several critical predictors of post-discharge complications. A similar pipeline can be adapted to different surgical centers, addressing specific population characteristics and resource limitations.

4.3.6 Real-World Validation

The development process was meticulously designed and implemented to ensure the robustness and trustworthiness of the DSS. As part of the CardioFollow.AI project, a clinical study was conducted at HSM, a unique opportunity to evaluate the impact of the RPM system. This study adhered to the principles of the Declaration of Helsinki and was approved by the hospital's Ethical Commission.

The study's objective extended beyond this thesis's scope, aiming to validate the impact of implementing an RPM system in clinical practice, as detailed in Section 4.2. The randomized control trial defined two standards of care: one group received the standard care protocol, while the other group also received an RPM kit.

To ensure similar populations in both groups while prioritizing higher-risk patients, the risk model was deployed weekly, predicting outcomes for all patients scheduled for discharge in the following days. The risk predictions were ranked, and the highest-risk patient was randomly assigned to one of the study groups. Subsequent patients were alternately assigned to each group in an intercalated manner.

Although the study's conclusions are still pending, the impact of the risk model's performance can be evaluated through extrapolation. Since all patients had their clinical information processed by the risk model, and the hospital's standard follow-up protocol already collects outcomes related to severe complications, ground truth labels for all patients included in the trial are available.

Table 4.9 presents the risk prediction results for the 228 surgeries performed between May 2022 and September 2023. With an average risk probability of 41.5%, the number of patients classified as high-risk exceeds those classified as low-risk. This imbalance is due to the threshold optimization strategy prioritizing sensitivity over specificity by a 2:1 ratio, resulting in a higher prediction of high-risk patients. Despite the imbalance, the model successfully identified all patients who subsequently experienced complications, as there were no false negatives.

Table 4.9: Description of model performance during the clinical study.

Surgeries	Average Score	No. Low Risk	No. High Risk	True Positives	False Positives	False Negatives
228	41.5%	39	189	11	142	0

Additionally, while the model was initially developed and validated using patient

data up to 2019, new registries have been collected throughout the project. These updated datasets provide an opportunity to revalidate the model's performance and assess its robustness over time. Table 4.10 presents the results obtained from the datasets collected during the real-world validation of the system. When compared to previous test periods, the AUROC for Test C (2020-2021) remains consistent with prior years, particularly aligning closely with Test A (2019). This suggests that the data from 2019 onward may exhibit subtle differences that account for the slight variation in performance. Moreover, the remaining metrics indicate that the current threshold results again in more high-risk predictions at the expense of accurate negative case detection, as evidenced by the lower specificity. However, since the model has maintained its predictive power, a new threshold optimization phase could be considered to rebalance the metrics if necessary.

Table 4.10: Final results validated against the test sets, in terms of AUROC, Sensitivity, and Specificity. The highest value for each metric and parameter is highlighted.

Dataset	AUROC (%)	Sensitivity (%)	Specificity (%)
Train (5-Fold CV)	71.7 $\pm$ 1.9	67.8 $\pm$ 9.3	66.4 $\pm$ 8.3
Test A (2019)	65.3	52.2	70.0
Test B (2008-2018)	69.5	66.0	65.2
Test C (2020-2021)	65.8	75.0	42.4

This system was designed to assist in the allocation of RPM resources, adding an extra layer of care without any risk of discrimination in the care procedure. Therefore, the results are impactful for clinical practice, as intended. Although the model's performance metrics may not appear exceptional (see Table 4.7), its real-world impact and robustness are evident.

4.3.7 Limitations

The proposed model for risk prediction in cardiothoracic surgery advances existing literature by predicting over a longer period and incorporating a broader set of variables for better RPM resource allocation. While it enhances the robustness of real-world models for clinical practice, several limitations remain.

Firstly, the clinical data was sourced from a single hospital, limiting the generalizability of the findings across different populations and settings. This may affect the model's performance when applied to diverse patient demographics and clinical environments. Expanding data collection to include multiple hospitals could enhance the model's generalizability, robustness, and potentially its performance.

Secondly, the issue of data imbalance remains partially unresolved, as reflected in the model's sensitivity, which required improvement through an unconventional threshold optimization ratio. Implementing advanced techniques such as synthetic data generation or enhanced sampling methods could better address data imbalance, improving both

sensitivity and overall model performance. However, the data imbalance problem, which is recurrent and critical in medical scenarios, remains largely unsolved [110].

Additionally, the model's performance is limited by the lack of important predictive factors, such as continuous clinical variables or external factors during home recovery. This could enhance the system's predictive accuracy and robustness. The absence of these variables might limit the model's ability to fully capture the complexity of patient risk profiles. RPM can thus become an important tool to provide data towards a more continuous framework for risk prediction.

Finally, the current algorithm is static and does not adapt to new data or evolving data distributions. This lack of adaptability could limit the model's long-term effectiveness and accuracy as new patient data becomes available or clinical practices change, as identified when the model was tested on new data. Transitioning to an incremental learning approach may allow the model to continuously update, ensuring sustained performance improvements and relevance. Nevertheless, concerns related to the loss of human supervision or the impracticability of constant supervision raise questions about the viability of medical DSSs.

This work addresses immediate needs and inspires future research to overcome these limitations and advance the field of risk prediction in cardiothoracic surgery.

4.4 Handling Missing Data from the Real World¹

*This has been joint work with Isabel Curioso,
conducted as part of her Master's project.*

4.4.1 Objectives and Significance

One of the key challenges in working with medical data is its inherent heterogeneous nature. Medical data often originates from multiple sources, spans diverse contexts, and exists in various formats, resulting in inconsistent and incomplete datasets. Missing data is prevalent in real-world settings, yet its impact on the robustness of DSSs is often overlooked [143]. Public datasets are often pre-processed and harmonized, allowing researchers to bypass the challenges of missing data. When researchers face the problem, they usually use simple imputation techniques. Traditional methods like deletion or mean-filling fail to account for the distinct sources of missingness explored in Section 2.3.1. The imputation process's quality directly impacts model performance, making it a crucial factor in building reliable and robust DSSs [87].

¹This Section is based on the joint work with Isabel Curioso while guiding her Master's project, which culminated in the publication "Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-based Imputation" by Isabel Curioso, **Ricardo Santos**, Bruno Ribeiro, André V. Carreiro, Pedro Coelho, José Fragata, and Hugo Gamboa, published in the *Journal of King Saud University - Computer and Information Sciences* (2023) [72].

Previous efforts, such as those detailed in Section 4.3, involved collaboration with clinicians to adopt a human-centric approach, improving data processing and informed imputation [280]. However, this is not always feasible. Domain-specific knowledge is sometimes limited, and the scale of datasets and the number of variables often make manual efforts impractical. In such cases, generalizable automated techniques become essential for effectively handling different types of missing data.

The literature has explored correlation as a basis for developing new techniques for missing data imputation, even though correlation does not imply causation. In clinical datasets, different features often represent various manifestations of the same physiological process or medical condition. This dependency between features creates an opportunity to leverage correlation as a powerful tool for imputing missing data. However, the potential of correlation-based imputation has not yet been fully explored in the context of health data, particularly in cardiothoracic surgery.

Given the limitations of previous work, **this study aims to propose and evaluate the impact of three novel correlation-based imputation techniques in scenarios where domain knowledge is limited or datasets are too large for manual curation.** The proposed methods explore not only the correlation between values but also the relationship between values and missingness patterns. Although the primary focus is on the challenges of the cardiothoracic surgery dataset, three additional public datasets are utilized to develop and validate the proposed methodologies. These data-centric strategies lay the groundwork for more robust and automated approaches to addressing the challenges of heterogeneous and incomplete medical datasets.

4.4.2 Literature Review

Addressing missing data during data processing is required for ML models that expect a clean, structured, and harmonized input matrix.

Deletion methods, which involve excluding incomplete data, are simple and convenient, as demonstrated by Zhou et al. [363]. However, such techniques can introduce bias into the analysis [88]. Imputation methods, on the other hand, replace missing values with predictions based on observed data. Standard statistical and ML-based techniques include mean imputation, regression imputation, stochastic regression imputation, hot-deck imputation, and KNN imputation [88]. Multiple imputation approaches, such as the Multivariate Imputation by Chained Equations (MICE) algorithm developed by van Buuren and Groothuis-Oudshoorn [52], have also gained traction.

Data-splitting-based techniques have gained popularity for handling missing data by dividing the dataset into subsets or classes before performing imputation. One notable method is Class Center-based Missing Value Imputation (CCMVI) [321], which partitions the dataset into classes and computes class centers, standard deviations, and thresholds based on complete data samples for imputing missing values. While CCMVI shows improved performance over other methods, specifically for MCAR, it assumes that missing

data only occurs in the training set, which is often impractical. To address this, Bhagat and Singh [39] proposed the Nullify the Missing Values before Imputation (NMVI) method, which eliminates the need for a complete data subset by replacing missing values with zero and utilizing all class samples for imputation. NMVI showed comparable performance to other imputation techniques, but it still leaves questions about handling missing data in the test set.

4.4.2.1 Correlation-based Imputation

Recently, correlation-based imputation techniques have been introduced. Mishra et al. [219] proposed Feature Correlation based Missing Data Imputation (FCMI). This method replaces missing values in an attribute by using regression models trained on features that are highly correlated with the incomplete attribute. FCMI was evaluated on the MAR missingness mechanism with distinct missing rates but depends on a complete subset of samples to compute the correlations.

Similarly, Sefidian and Daneshpour [285] developed a framework with ten regression-based algorithms called Correlation Maximization-based Imputation Methods (CMIM), which maximize the correlation between missing and observed attributes. The authors hypothesized that linear correlations are stronger within specific subsets of the data than across the entire dataset, leading them to segment the data into highly correlated groups before fitting linear regression models for imputation. Their approach, evaluated under the MAR mechanism with multiple missing rates, demonstrated superior performance compared to traditional methods. However, a fundamental limitation of CMIM is the need for a manual selection of the best technique for each dataset, as no single method consistently outperformed the others, making the process both time-consuming and non-automatic.

Building on the popularity of KNN-based imputation methods, Liu, Lai, and Zhang [196] developed the Correlation-based Hierarchical K-Nearest Neighbors (CoHiKNN) algorithm. CoHiKNN leverages correlations between attributes as weights in calculating distances between complete and target incomplete records. After selecting the nearest neighbors, the missing values are imputed using the average of the observed values in a phased way, starting with records that have the fewest missing values and progressively including records with more missing data. Initially, correlation coefficients are derived through a complete-case analysis and are updated as imputed data is incorporated. Despite its strong performance compared to traditional methods, including standard KNN imputation, CoHiKNN requires an initial complete subset representative of the whole dataset for calculating the first set of correlation coefficients. This requirement may be difficult to meet with real-world datasets.

Khan et al. [164] further advanced this concept with the Convolutional Neural Network Imputation (CNNI) approach, which identifies correlations within the dataset to train a convolutional kernel that replaces missing values. CNNI demonstrated comparable or

superior performance to traditional methods. Still, experiments were conducted solely on numerical data under a MCAR mechanism, raising questions about the generalizability of the results to more complex real-world scenarios.

In the context of real-world clinical datasets, Yoon, Zame, and van der Schaar [353] developed a DL model that exploits feature relationships, particularly correlations within and across data streams, to impute missing values. While Yoon’s work was not tied to a specific clinical case study, the model was tested on multiple real-world medical datasets and demonstrated significant improvements over state-of-the-art techniques, including MICE and other DL-based methods. Regarding Alzheimer’s disease research, Tabarestani et al. [309] developed a multitask learning method that handles missing data by projecting high-dimensional input features into multiple low-dimensional, less-sparse spaces. This approach was designed to be robust to the adverse effects of missing data by capturing dependencies between different feature representations, ensuring a relatively high accuracy despite incomplete datasets.

Despite notable progress in addressing missing data, many state-of-the-art imputation techniques exhibit significant limitations, particularly in real-world clinical settings. One of the most common strategies remains complete-case analysis, disregarding the potential informational value of missingness itself. Research suggests that missing data can be informative, and inattentive complete-case analysis may lead to biased results and compromised model performance since a complete and representative subset of data is rarely feasible in healthcare [196, 219, 285].

Validation of these techniques often focuses on numeric variables alone, neglecting the categorical and mixed data types prevalent in medical datasets. Although correlation has been shown to improve imputation, its potential remains underexplored in medical research. Most ML applications view highly correlated features as redundant, discarding them prematurely, even though recent studies highlight the advantages of incorporating correlation into imputation strategies. Given the inherent variability of real-world clinical data, further exploration of correlation-based methods could lead to more robust and interpretable imputation techniques better suited to healthcare applications.

Despite the significant research efforts toward using correlation to enhance the robustness of DSSs through effective missing value imputation, existing strategies still face notable limitations when applied to real-world data. Most approaches address only a single type of missingness mechanism, despite real-world datasets often containing multiple types of missingness, which can be difficult to identify accurately. Additionally, many imputation methods rely on the availability of a complete subset of data to compute correlations, a requirement rarely met in practice.

4.4.3 Methodology

This work introduces three novel imputation techniques designed to address these shortcomings. By leveraging correlation in a more flexible and innovative manner, these

methods aim to improve the accuracy and robustness of imputation in clinical datasets, providing more reliable support for real-world healthcare applications.

The proposed imputation techniques explore the potential advantages of incorporating correlations between values and missingness patterns instead of relying solely on the correlation between existing values like previous works have done. The underlying hypothesis is that if a strong correlation exists between the two, the missing data in one variable may be concentrated within a specific range of the distribution of another variable's values. Recognizing this association could provide valuable insights to estimate the missing elements accurately.

4.4.3.1 Datasets

This work builds on the theme of this chapter, aiming to develop tools that enhance the robustness of DSSs in cardiothoracic surgery. For this purpose, we leveraged the dataset described in Section 4.3.3.1, which includes all patients who underwent cardiothoracic surgery at HSM prior to clinicians-informed data processing. The task is to predict the risk of severe complications within 90 days post-hospital discharge. The dataset consists of 5,625 samples with a positive label rate of 7.3%. The initial set of 105 variables, along with two additional variables aggregating hospital complications, were used for model development. Table 4.11 provides a detailed summary of the average missing rate identified in these variables.

UCI Machine Learning Repository Datasets

Given that the cardiothoracic surgery dataset contains missing data from the source, we also used three complete datasets from the University of California Irvine (UCI) Machine Learning Repository [160] to evaluate the developed models under various synthetic missingness ratios:

- **Wine dataset:** This dataset contains 178 samples resulting from the chemical analysis of three types of wine, classified into three categories (wine types) in a multi-class objective. It includes 13 numerical attributes.
- **SPECT Heart dataset:** Derived from feature extraction of Single Proton Emission Computed Tomography (SPECT) images, this dataset classifies samples into normal or abnormal categories in a binary classification task. It consists of 187 patient samples and 22 binary attributes.
- **Statlog Heart dataset:** This dataset assesses the presence of heart disease using 13 numerical and categorical variables related to cardiac conditions, with data from 270 patients.

Table 4.11: Summary of missing rates across different variable groups in the cardiothoracic surgery dataset, expressed as mean percentages with standard deviations.

Feature Group	Attributes	Average Missing Rate (%)
Hospitalization	4 dates	23.7 ± 47.2
Cardiac history	4 ordinal, 1 binary	16.1 ± 35.3
Previous interventions	2 dates, 1 ordinal, 1 nominal	44.9 ± 51.6
Pre-operative risk factors	4 numeric, 3 ordinal, 4 nominal, 4 binary	0.9 ± 0.2
Pre-operative haemodynamics and catheterization	1 date, 5 numeric, 2 ordinal, 1 nominal, 1 binary	49.3 ± 42.8
Pre-operative status and support	4 binary	0.6 ± 0.2
Operation	1 text, 2 ordinal, 4 nominal	27.2 ± 46.1
Coronary surgery	2 numeric, 2 nominal	61.4 ± 0.7
Valve surgery	1 code, 8 numeric, 3 ordinal, 10 nominal, 7 binary	34.8 ± 30.2
Cardiac Surgery Morbidity Scale	1 numeric, 8 nominal	7.6 ± 22.1
Discharge details	2 nominal, 1 binary	32.4 ± 55.6
Patient demographics and autocalculations	1 code, 10 numeric, 1 nominal, 1 binary	7.8 ± 26.4

These datasets were chosen to represent different types of data. Multi-class nominal variables were one-hot encoded, and ordinal encoding was applied to ordinal attributes, transforming all features into either numeric or binary formats.

To evaluate the performance of the proposed imputation methods, synthetic missing values were introduced into the three datasets under all three missingness mechanisms, with missing rates of 10%, 30%, and 50% applied to each attribute. All features were made incomplete for the MCAR mechanism and shared the same missing rate. In contrast, for the MAR and MNAR mechanisms, only 50% of the features were incomplete, but they maintained a consistent missing rate. This resulted in nine synthetic datasets for each UCI Machine Learning Repository dataset. The R package `missMethods` [269] injected the missing data, allowing for a comprehensive study that explored various missing rates and mechanisms.

A medical dataset related to osteoporosis was also utilized for further evaluation. However, its description, results, and specific discussion are detailed in the original publication [72].

4.4.3.2 Notation

The proposed methods share common principles regarding the use of correlation for missing data imputation. Therefore, we can establish a general notation that will be used throughout.

Consider a dataset with a total of G features, where $G_{\text{miss}} \leq G$ denotes the number of features that contain missing values. Let g_j represent the j th feature, where $j \in \{1, 2, \dots, G\}$, and $g_{\text{miss},i}$ represent the i th incomplete feature, where $i \in \{1, 2, \dots, G_{\text{miss}}\}$. Additionally, define C_{vm} as a $(G_{\text{miss}} \times G)$ matrix. For each pair $\{i, j\}$, where $g_{\text{miss},i} \neq g_j$, the element $C_{\text{vm}}[i, j]$ stores the correlation between the values of g_j and the missingness pattern of $g_{\text{miss},i}$, represented as a binary indicator of whether the data is missing.

Furthermore, let C_{vv} represent the standard $(G \times G)$ correlation matrix. This matrix is computed using a pairwise deletion strategy, meaning the correlation between each pair of features is calculated based only on the available data on an analysis-by-analysis basis. Figure 4.6 schematizes C_{vv} and C_{vm} .

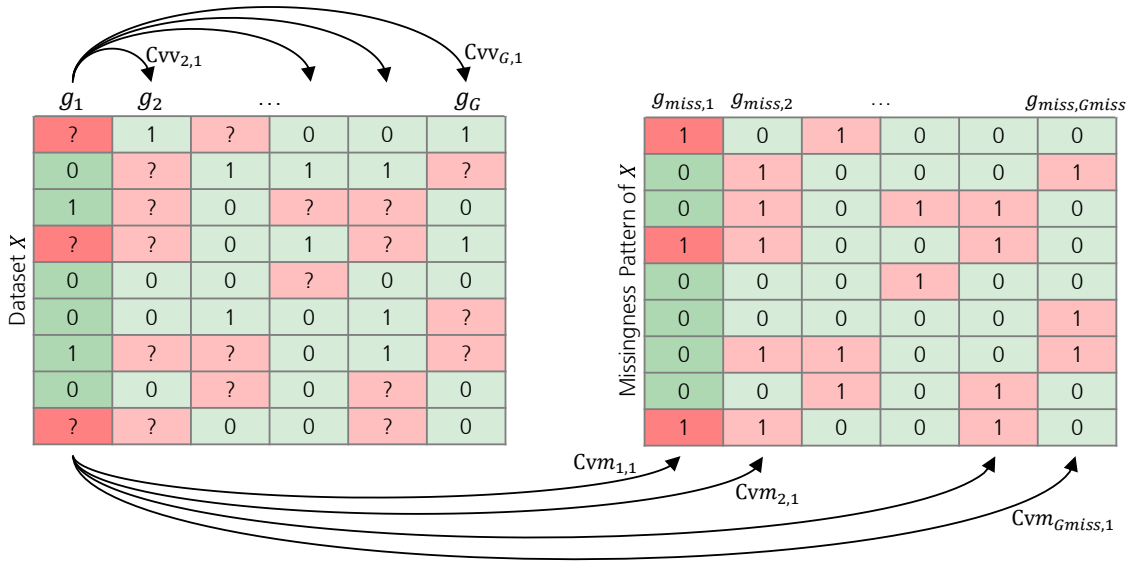


Figure 4.6: Depiction of the two types of correlation computed. This example only looks at the correlation values for the first feature, i.e., $g = 1$.

4.4.3.3 Correlation Weighted K-Nearest Neighbors Imputation

The Correlation Weighted K-Nearest Neighbors Imputation (CWKNNI) method draws inspiration from the CoHiKNN technique proposed by [196]. However, CWKNNI improves upon it by not only considering the correlation between attribute values but also incorporating a weighted average that includes the correlation between values and missingness patterns. Furthermore, CWKNNI addresses the limitation of CoHiKNN's reliance on a complete data subset for imputation. Instead, it calculates the correlation matrix using a pairwise deletion approach, avoiding the need for list-wise deletion.

The process for implementing CWKNNI is depicted in the flowchart of Figure A.1a and outlined in the following step-by-step explanation:

1. Consider a dataset X , with G features and N instances.
2. Compute the C_{vv} and the C_{vm} correlation matrices. This approach considers the absolute value of these correlations, thus accounting for the strength of the association, not its direction.
3. Order the features from lowest to highest missing rate. Imputation will be performed in this sequence in a phased manner.
4. Let g_j denote the attribute being imputed. Create a subset X_{miss_j} consisting of samples where g_j is missing. Furthermore, create a subset X_{obs_j} with samples where g_j is observed. If X_{miss_j} is empty, skip to the next attribute.
5. For each instance in X_{miss_j} find the k nearest neighbors within X_{obs_j} . A weighted Euclidean distance, which accounts for the presence of missing values, is used as a distance measure:

$$d_{vt} = \sqrt{w_D \times \sum_{h \in O_g} (1 - w_{Ch}) \times (v_h - t_h)^2} \quad (4.1)$$

where d_{vt} is the distance between samples v and t , v is an instance of X_{miss_j} , t is a sample from X_{obs_j} . v_h and t_h are the observed values from the h th attribute of v and t respectively. O_g is the set of indexes of the variables that are not missing in v nor t . w_D is the quotient between the total number of features g and the dimension of O_g . As for w_{Ch} , it is a weighted average of the correlation between g_j and g_h , and the correlation between the values of g_h and the missingness pattern of g_j :

$$w_{Ch} = p \times C_{vv}[h, j] + (1 - p) \times C_{vm}[h, j], \quad p \in [0, 1] \quad (4.2)$$

Note that $0 \leq w_{Ch} \leq 1$.

6. Replace the missing value of g_j in each instance of X_{miss_j} with a weighted prediction, in which the weights are the inverse of the computed distances d_{vt} : a closer neighbor has higher importance (i.e., weight) in the final prediction. The mean value is used for numeric variables, whereas the mode is used to impute binary attributes.
7. Repeat Steps 4-6 until all missing values from all features have been imputed.

The CWKNNI method involves two key parameters: the number of neighbors k and the percentage p , which determines the weight given to the correlation between values compared to the correlation between values and the missingness pattern.

Additionally, the method includes a boolean parameter, *initial_fill*, which specifies whether an initial temporary imputation using the standard KNN method is performed.

If enabled, this imputed dataset is used to calculate both C_{vv} and C_{vm} . Furthermore, the subset X_{obs_j} will consist of these KNN-imputed samples. Although the number of samples in each X_{obs_j} remains unchanged, all instances will be complete without missing attributes. The only missing feature for the sample being imputed will be g_j .

Finally, the boolean parameter *update* controls whether the imputed values are used in subsequent phases of imputation rather than relying solely on the original values of each sample. At the start of each phase, both the C_{vv} and C_{vm} matrices are updated to reflect the newly imputed values. However, it is important to note that the missingness patterns used to calculate C_{vm} remain unchanged.

4.4.3.4 K-Nearest Neighbors Selected by Correlation Imputation

Like most KNN-based methods, CWKNNI has increased computational demands when dealing with high-dimensional data since distance must be calculated in each dimension. The K-Nearest Neighbors Selected by Correlation Imputation (KNNSCI) method was developed to address this limitation. KNNSCI reduces dimensionality by performing a pre-selection of features based on correlation for each variable to be imputed, making it more suitable for large datasets. Although both CWKNNI and KNNSCI are KNN-based approaches, KNNSCI operates differently by leveraging correlation in a distinct way. While CWKNNI incorporates correlation into a weighted distance function, KNNSCI uses a standard Euclidean distance. Additionally, KNNSCI employs correlation to select relevant features before imputation, a step that CWKNNI does not perform.

Figure A.1b includes the flowchart for the KNNSCI method, and the sequence of steps necessary to replicate the method are:

- 1-4. Refer to steps 1 to 4 from CWKNNI.
5. Using the correlation matrices, compute the coefficient $r_{Cj,h}$ between attribute g_j and the h th attribute of X (apart from g_j). This coefficient is equal for all samples within the subset X_{miss_j} and is obtained through the following equation:

$$r_{Cj,h} = p \times C_{vv}[h, j] + (1 - p) \times C_{vm}[h, j], \quad p \in [0, 1] \quad (4.3)$$

Note that $0 \leq r_{Cj,h} \leq 1$.

6. Create a subset X_{obs_j} with samples where g_j is observed. Find the features where $r_{Cj,h} < threshold_c$ and remove their columns from both X_{miss_j} and X_{obs_j} . The $threshold_c$ parameters should only select highly correlated features. Based on the work of Schober et al. [284], $threshold_c$ was set at 0.7 since the authors state that a coefficient above 0.7 indicates a strong correlation.
7. For each instance in X_{miss_j} find the k nearest neighbors within X_{obs_j} (after removing the columns with $r_{Cj,h} < threshold_c$). A standard Euclidean distance is used as a distance measure.

8. Replace the missing value on the attribute g_j in each instance of X_{miss_j} with a weighted prediction, in which the weights are the inverse of the computed Euclidean distances: a closer neighbor has a higher importance in the final prediction. The mean value is used for numeric variables, whereas the mode is used to impute binary attributes.
9. Repeat Steps 4-8 until all missing values from all features have been imputed.

The parameters necessary to KNNSCI are similar to CWKNNI, namely the number of neighbors k and the percentage p , as well as *initial_fill* and *update*.

4.4.3.5 Correlation Weighted Regression Imputation

The Correlation Weighted Regression Imputation (CWRI) method is a regression-based approach where missing values are imputed using predictions from multiple linear regression models. For each missing value, several estimates are generated and then combined, with the final imputation determined by the correlational significance of the predictor variables. This correlational importance is calculated as a weighted average, incorporating the correlation between values and between values and missingness patterns.

The flowchart of CWRI is displayed in Figure A.2 and the step-by-step implementation is as follows:

1. Consider a dataset X , with G features and N instances. G_{miss} is the number of attributes with missing values.
2. Compute C_{vv} and C_{vm} , with the correlations between values and missingness patterns. This approach leverages the absolute value of these correlations.
3. Build a $g \times g$ matrix, denoted as R , containing multiple linear regression models. Let g_h and g_j be two different attributes of X , with indexes h and j , respectively. For every pair $\{h, j\}$, where $h \neq j$, $R[h, j]$ stores a linear regression model in which g_h is the independent variable and g_j is the predictor.
4. Assemble the subset X_{inc} including all incomplete samples, i.e., instances with at least one missing value. N_{inc} is the total number of incomplete samples.
5. Consider any instance of X_{inc} , and let g_i denote its i th missing attribute. Using every non-missing feature g_t in this sample, obtain the corresponding predicted value for g_i through the regression model $R[i, t]$. If g_i is a numeric variable, the final imputation will be a weighted average of all predicted values. If g_i is binary, a weighted mode is applied. To obtain the weight given to each variable g_t , denoted as $w_{i,t}$, first calculate its correlational importance $w_{Ci,t}$ through the following equation:

$$w_{Ci,t} = p \times C_{vv}[t, i] + (1 - p) \times C_{vm}[t, i], \quad p \in [0, 1] \quad (4.4)$$

Note that $0 \leq w_{Ci,t} \leq 1$. After computing $w_{Ci,t}$ for every non-missing feature, $w_{i,t}$ is a simple normalization:

$$w_{i,t} = \frac{1}{\sum_t w_{Ci,t}} \times w_{Ci,t} \quad (4.5)$$

6. Repeat Step 5 until all incomplete samples have been imputed.

The only parameter in CWRI is the percentage p , which determines the weight given to the correlation between values relative to the correlation between values and missingness patterns.

4.4.4 Experiments

A comprehensive experimental study was conducted to evaluate the performance of the proposed imputation techniques. This study involved testing against seven other widely used imputation methods: Mean/Mode imputation, Regression, KNN, CMIM [285], CoHiKNN [196], NMVI [39], and MICE [52]. Additionally, list-wise deletion performance was included in the evaluation for comparison. The selection of these methods aimed to cover both standard and more recent techniques, ensuring a broad range of baseline strategies.

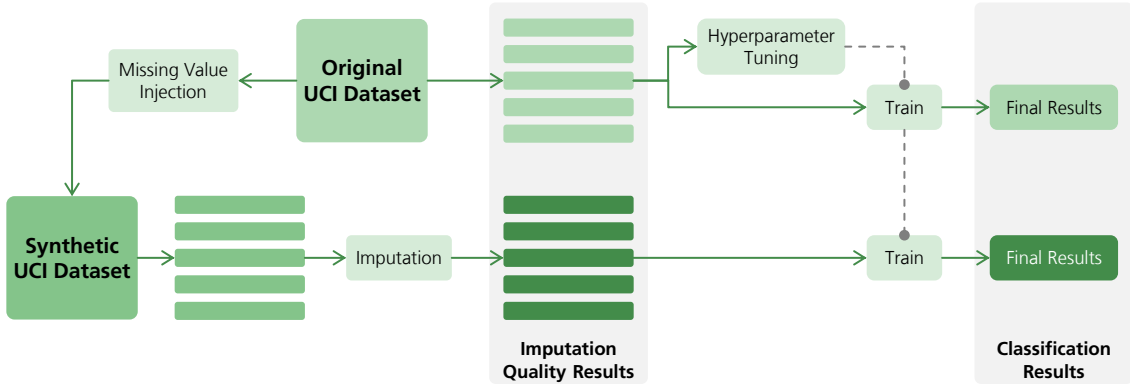
Before imputation, a stratified 5-fold CV strategy was applied to the UCI Machine Learning Repository datasets to maintain the positive rate among train and test sets. Additionally, this split was performed after the injection of synthetic missing values discussed in Section 4.4.3.1. Since the imputation methods use metrics that depend on the full set, as in a real-world scenario, the missingness patterns are expected to not change between data partitions.

Regarding the cardiothoracic surgery dataset, a group stratified 5-fold CV strategy was utilized to guarantee no data leakage of samples from the same patient in different partitions. Additionally, the test set with all data from 2019 described in Section 4.3.3.2 was kept separate and used to evaluate the different mechanisms. The evaluation was conducted in terms of imputation quality and classification performance. Figure 4.7 schematizes the evaluation pipelines implemented for the different datasets.

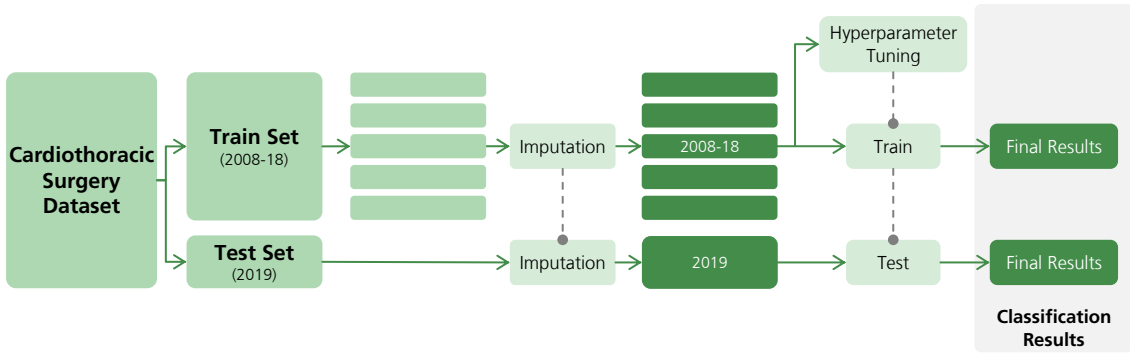
4.4.4.1 Imputation Quality

The first type of evaluation focused on assessing the accuracy of the imputation procedure by comparing the imputed values to the original values, i.e., the ground truth. This evaluation was carried out on the incomplete datasets generated from the three UCI Machine Learning Repository datasets since their actual values were known before synthetic missingness was introduced. The MAE was used as the primary metric for measuring imputation quality. The schematic of this evaluation is pictured in Figure 4.7a.

Each imputation method's performance was rigorously compared to determine the effectiveness of the proposed techniques against existing methods. Since each imputation



(a) UCI Machine Learning Repository datasets



(b) Cardiothoracic Surgery dataset

Figure 4.7: Evaluation pipelines of the missing data imputation models in different datasets.

method presents distinct hyperparameters that can be adjusted, an initial hyperparameter tuning was performed to ensure that each method was evaluated under optimal conditions. Table A.3 details the range of parameters tested for each method.

4.4.4.2 Classification Performance

The second type of evaluation examined the impact of different imputation techniques on ML model performance. Three classifiers were tested: RF, SVM, and NB. A 5-fold CV strategy was applied, and the AUROC was used as the primary metric for comparing classifier performance across the various imputation methods.

For the UCI datasets, a grid search was conducted on the original data to determine the optimal hyperparameters for each classifier, as depicted in Figure 4.7a. However, for the cardiothoracic surgery datasets, where ground truth for the missing values was unavailable, hyperparameter tuning was performed on each imputed dataset to ensure a fair comparison, as schematized in Figure 4.7b. The computational cost related to the larger dataset size was mitigated by reducing the tested hyperparameter ranges for the RF classifier. Table A.4 lists the hyperparameters tested for each dataset type. The classifier with the highest average AUROC across the folds was considered the best-performing

model. No threshold optimization was performed, and the default value of 0.5 was used for converting predicted probabilities into class labels.

This rigorous and systematic approach ensured that the comparative study was as fair and accurate as possible, minimizing the influence of factors unrelated to the imputation process.

4.4.5 Results and Discussion

Experiments tested a wide range of parameters for both imputation methods and classification models. The best combinations between model and imputation method for every dataset are detailed in Table A.5 of Appendix A. Experimental results regarding imputation quality and classification performance are presented and discussed throughout.

4.4.5.1 Imputation Quality

Given the large volume of results and the goal of conducting a comprehensive comparative study, it was necessary to aggregate them. Precisely, the average of the three MAEs corresponding to the different missing rates (10%, 30%, and 50%) was calculated for each missingness mechanism. While this approach does not allow for an analysis of the impact of missing rates on the imputation process, evaluating the missingness mechanism's overall effect was considered more important.

The results for each UCI dataset are displayed in Table 4.12. Techniques that could not perform imputation at higher missing rates are marked, as the reported errors do not fully reflect their overall efficacy. Since MAE generally increases with higher missing rates due to reduced available information, the actual performance of these techniques may be worse than indicated by their MAE values. Consequently, while marked techniques sometimes show the lowest MAEs, they cannot be considered the most effective in imputation quality.

The proposed correlation-based imputation methods consistently deliver strong performance across all three missingness mechanisms. In particular, CWKNNI stands out as the most accurate technique under the MCAR mechanism. Furthermore, the ranking of imputation quality remains relatively stable across different datasets. For example, techniques that perform well on the Wine dataset show the same outcome on the SPECT Heart and Statlog Heart datasets. This consistent ranking suggests that CWKNNI produces estimates closer to the actual values in MCAR data, regardless of whether the data types are numeric or binary.

A distinct tendency is verified when the proposed methods are applied to the MAR mechanism, as there is no clear superiority. Since the injection of MAR values was based on pairs of highly correlated features, where one feature determines the missing values in the other, the proposed methods were expected to outperform other techniques due to their use of correlation in the imputation process. Although their performance was among the best, it did not demonstrate a clear advantage. The diminishing correlation

Table 4.12: Results of missing imputation quality for all missingness mechanisms, as the mean and standard deviation of MAE. Techniques that failed at higher missing rates were never selected as the best since their MAEs do not fully capture their overall performance.

Imputation Method	Wine	SPECT Heart	Statlog Heart
<i>MCAR mechanism</i>			
Mean / Mode	0.18 ± 0.00	0.37 ± 0.01	0.25 ± 0.01
KNN	0.14 ± 0.01	0.26 ± 0.04	0.17 ± 0.05
NMVI	0.18 ± 0.03	0.42 ± 0.02	0.29 ± 0.00
CMIM	0.15 ^(a)	0.29 ^(a)	0.18 ^(a)
CoHiKNN	0.15 ^(a)	0.29 ^(a)	0.20 ^(a)
Regression	0.15 ^(a)	0.29 ^(a)	0.19 ^(a)
MICE	0.17 ± 0.01	0.34 ± 0.05	0.23 ± 0.04
CWKNNI*	0.13 ± 0.01	0.25 ± 0.03	0.16 ± 0.03
KNNSCI*	0.13 ± 0.01	0.26 ± 0.04	0.17 ± 0.03
CWRI*	0.15 ± 0.00	0.29 ± 0.01	0.18 ± 0.02
<i>MAR mechanism</i>			
Mean / Mode	0.23 ± 0.01	0.46 ± 0.05	0.35 ± 0.08
KNN	0.13 ± 0.03	0.19 ± 0.09	0.21 ± 0.11
NMVI	0.14 ± 0.02	0.43 ± 0.06	0.32 ± 0.12
CMIM	0.19 ± 0.04	0.23 ± 0.03 ^(b)	0.21 ± 0.08
CoHiKNN	0.19 ± 0.02	0.16 ± 0.01 ^(b)	0.28 ± 0.10
Regression	0.19 ± 0.05	0.13 ± 0.01 ^(b)	19.02 ± 20.38
MICE	0.15 ± 0.03	0.29 ± 0.05	0.18 ± 0.03
CWKNNI*	0.13 ± 0.03	0.22 ± 0.10	0.24 ± 0.11
KNNSCI*	0.13 ± 0.03	0.20 ± 0.09	0.23 ± 0.11
CWRI*	0.18 ± 0.02	0.25 ± 0.13	0.27 ± 0.07
<i>MNAR mechanism</i>			
Mean / Mode	0.35 ± 0.00	0.62 ± 0.11	0.41 ± 0.13
KNN	0.24 ± 0.05	0.38 ± 0.17	0.30 ± 0.11
NMVI	0.16 ± 0.04	0.49 ± 0.08	0.33 ± 0.05
CMIM	0.27 ± 0.06	0.41 ± 0.18	0.18 ± 0.02 ^(b)
CoHiKNN	0.29 ± 0.02	0.41 ± 0.27	0.30 ± 0.04 ^(b)
Regression	0.27 ± 0.07	0.24 ± 0.04 ^(b)	0.20 ± 0.01 ^(b)
MICE	0.24 ± 0.05	0.35 ± 0.08	0.27 ± 0.03
CWKNNI*	0.24 ± 0.05	0.39 ± 0.20	0.31 ± 0.12
KNNSCI*	0.24 ± 0.05	0.39 ± 0.16	0.30 ± 0.11
CWRI*	0.31 ± 0.01	0.45 ± 0.22	0.36 ± 0.07

^(a) The technique could not perform imputation for the missing rates of 30% and 50%.

^(b) The technique could not perform imputation for a missing rate of 50%.

* Proposed algorithms.

between values, which becomes more pronounced as the missing rate increases, may have impacted the precision of the imputation.

The MNAR mechanism generally produced the highest MAEs across all imputation methods, as expected, given that most traditional or state-of-the-art techniques are primarily designed for MAR scenarios. As a result, these methods perform better under this and MCAR mechanisms. The proposed methods also exhibited higher MAEs in the MNAR mechanism, with performance comparable to that of other imputation techniques.

When considering the use of correlation for missing value prediction, the proposed methods CWKNNI and KNNSCI demonstrated overall superior performance compared to other correlation-based methods, such as CMIM and CoHiKNN. Additionally, unlike the competing methods, CWKNNI and KNNSCI could handle imputation across all missing rates, further highlighting their robustness and effectiveness.

4.4.5.2 Classification Performance

Table 4.13 presents the results of the classification performance evaluation, where three ML classifiers (RF, SVM and NB) were trained on the imputed datasets and compared based on their AUROC. For the synthetic datasets, the AUROCs were averaged across all missing rates for each missingness mechanism. Some techniques could not perform imputation for specific datasets due to their restrictive requirements, as marked in the Table. The emboldened values represent the tests that led to the highest scores, although differences between techniques are usually not statistically significant. Furthermore, methods with incomplete tests were not selected as the best even though the available average is the highest. This decision is based on the fact that high missing rates are expected to negatively impact results since there is a more significant deviation from the actual data distribution.

The proposed correlation-based imputation methods generally performed on par with or even outperformed their competitors in several cases. Notably, RF models trained on datasets imputed using the KNNSCI method demonstrated the best overall performance, achieving the highest AUROCs in three evaluations, including the two cardiothoracic surgery sets. The combination of KNNSCI and RF led to the best results for the real-world clinical data, although not statistically significant, which is a crucial outcome when developing robust ML-based systems for deployment in healthcare environments.

Correlation-based imputation methods outperformed others in seven evaluations, with KNNSCI and CWRI emerging as the primary contributors to these strong results. In addition to producing higher AUROCs, these methods consistently performed well across all missing rates and mechanisms. The MICE imputation method appears as the leading competitor, which confirms its relevance in the literature. Nonetheless, its best results are attained in the NB classifier, which does not achieve the highest overall AUROCs.

The results also highlight the significant difference in AUROC scores produced by list-wise deletion compared to other methods. While simple to implement, list-wise deletion can lead to substantial losses in dataset representativeness, particularly at higher

4.4. HANDLING MISSING DATA FROM THE REAL WORLD

Table 4.13: Results of classification performance for all classifiers, as the mean and standard deviation of AUROC, in percentage (%). The emboldened values are each assessment's higher AUROCs. Techniques that failed in higher missing rates were never selected as the best result since the presented values do not capture their overall performance.

	Wine	SPECT Heart	Statlog Heart	Cardiothoracic Surgery (2008-2018)	Cardiothoracic Surgery (2019)
<i>RF classifier</i>					
Complete dataset	100.00 ± 0.00	81.04 ± 10.47	91.47 ± 3.90	N.A.	
List-wise deletion	79.70 ± 24.26 ^(a)	70.04 ± 11.94 ^(a)	91.38 ± 1.49 ^(b)	77.55 ± 14.35 ^(d)	56.13 ± 7.31
Mean / Mode	98.71 ± 0.94	83.77 ± 8.26	89.18 ± 2.22	70.18 ± 4.66	61.24 ± 0.97
KNN	98.45 ± 1.07	81.12 ± 7.12	89.05 ± 2.62	71.08 ± 4.94	61.17 ± 1.81
NMVI	98.27 ± 2.05	83.86 ± 6.03	88.06 ± 2.59	65.61 ± 5.04	57.07 ± 0.94
CMIM	99.24 ± 0.39 ^(a)	86.50 ± 5.50 ^(a)	89.23 ± 1.68 ^(b)	(c)	(c)
CoHiKNN	98.85 ± 0.51 ^(a)	88.93 ± 7.27 ^(a)	89.30 ± 1.44 ^(b)	(c)	(c)
Regression	98.90 ± 0.50 ^(a)	81.91 ± 5.03 ^(a)	88.11 ± 3.22 ^(b)	70.36 ± 4.81	60.16 ± 0.97
MICE	98.99 ± 0.69	83.02 ± 4.83	89.12 ± 1.96	70.63 ± 4.51	59.60 ± 1.41
CWKNNI*	98.56 ± 0.91	82.25 ± 5.19	89.33 ± 1.71	(c)	(c)
KNNSCI*	98.57 ± 0.98	84.63 ± 5.58	89.09 ± 2.07	71.32 ± 4.78	61.86 ± 1.00
CWRI*	98.72 ± 0.56	83.12 ± 6.26	89.50 ± 2.26	71.10 ± 1.92	60.23 ± 1.39
<i>SVM classifier</i>					
Complete Dataset	100.00 ± 0.00	83.39 ± 9.43	91.17 ± 3.17	N.A.	
List-wise Deletion	91.97 ± 15.99 ^(a)	58.39 ± 24.53 ^(a)	89.55 ± 2.9 ^(b)	46.37 ± 18.97	49.00 ± 8.66
Mean / Mode	99.27 ± 0.60	75.71 ± 24.30	88.37 ± 1.95	68.58 ± 2.89	55.95 ± 2.55
KNN	99.27 ± 0.62	79.44 ± 13.16	88.27 ± 2.32	68.88 ± 2.70	55.97 ± 2.44
NMVI	98.78 ± 1.72	84.72 ± 4.36	88.31 ± 2.09	68.54 ± 2.92	55.61 ± 2.56
CMIM	99.54 ± 0.34 ^(a)	85.42 ± 3.88 ^(a)	88.81 ± 1.35 ^(b)	(c)	(c)
CoHiKNN	99.50 ± 0.36 ^(a)	85.47 ± 5.84 ^(a)	89.26 ± 1.09 ^(b)	(c)	(c)
Regression	99.41 ± 0.49 ^(a)	83.85 ± 3.68 ^(a)	87.39 ± 3.90 ^(b)	63.72 ± 4.18	56.67 ± 1.64
MICE	99.35 ± 0.54	80.77 ± 14.89	88.49 ± 2.00	68.96 ± 2.54	55.56 ± 2.89
CWKNNI*	99.35 ± 0.50	80.91 ± 9.92	88.54 ± 1.46	(c)	(c)
KNNSCI*	99.23 ± 0.59	83.35 ± 6.00	88.35 ± 2.45	68.91 ± 2.68	56.14 ± 2.63
CWRI*	99.40 ± 0.44	84.32 ± 5.76	88.91 ± 1.84	68.58 ± 1.85	55.93 ± 3.24
<i>NB classifier</i>					
Complete Dataset	99.82 ± 0.25	74.86 ± 4.86	86.50 ± 4.33	N.A.	
List-wise Deletion	82.30 ± 20.3 ^(a)	68.97 ± 9.65 ^(a)	82.12 ± 6.69 ^(b)	50.85 ± 14.71	49.58 ± 3.16
Mean / Mode	97.79 ± 1.90	74.63 ± 5.95	85.50 ± 1.55	66.58 ± 4.57	58.00 ± 0.88
KNN	98.21 ± 1.13	75.92 ± 4.04	85.93 ± 1.99	67.05 ± 4.27	57.86 ± 0.78
NMVI	97.57 ± 2.21	74.32 ± 3.55	86.07 ± 1.81	66.90 ± 4.36	58.00 ± 0.85
CMIM	98.07 ± 1.48 ^(a)	77.97 ± 5.04 ^(a)	85.32 ± 1.42 ^(b)	(c)	(c)
CoHiKNN	98.58 ± 0.81 ^(a)	78.01 ± 5.85 ^(a)	85.84 ± 0.61 ^(b)	(c)	(c)
Regression	98.00 ± 1.50 ^(a)	74.45 ± 3.16 ^(a)	85.69 ± 2.37 ^(b)	50.80 ± 7.18	49.56 ± 3.97
MICE	98.45 ± 1.04	75.44 ± 2.72	87.38 ± 1.66	67.45 ± 4.07	58.04 ± 1.15
CWKNNI*	98.28 ± 1.08	77.36 ± 5.70	86.63 ± 0.91	(c)	(c)
KNNSCI*	98.30 ± 1.15	76.60 ± 3.74	86.19 ± 1.46	67.06 ± 4.26	57.94 ± 0.79
CWRI*	98.12 ± 1.40	75.11 ± 5.13	86.28 ± 1.57	65.07 ± 3.50	57.56 ± 0.68

(a) The technique could not perform under MCAR with missing rates of 30% and 50%.

(b) In addition to (a), the technique could not perform under MNAR with a 50% missing rate.

(c) The technique could not perform imputation due to high computational costs.

(d) Complete-case analysis only used 7.3% of the samples from the cardiothoracic surgery dataset.

* Proposed algorithms.

missing rates. This reinforces the need for caution when using such an approach, as it may inadvertently distort the original problem and undermine classification performance.

Interestingly, the assumption that better imputation accuracy directly translates to improved classification performance was challenged in this study. To explore this hypothesis, the classifiers were also trained on the original, complete UCI Machine Learning Repository datasets, which represent an ideal imputation scenario with no missing values (and hence a null MAE). Comparing the performance of these classifiers with those trained on imputed datasets allowed us to test whether better imputation quality leads to better classification outcomes. Additionally, it should be noted that hyperparameter tuning was performed on the original datasets during a grid search. In contrast, no such tuning was applied to classifiers trained on imputed datasets to optimize computational resources. This may have provided a slight performance advantage to the models trained on the original datasets.

However, in some cases, such as with the SPECT Heart dataset, RF models trained on imputed datasets outperformed those trained on the original dataset. This suggests that there is not always a direct relationship between imputation quality and ML model performance. This finding raises an important question: Should the ideal imputation method for ML-based clinical systems prioritize producing more accurate imputations, or should it focus on achieving better classification performance? While robust clinical DSSs should prioritize optimal performance, dismissing imputation quality entirely may not be advisable. For instance, an imputation method that introduces bias by distorting the original data distribution might improve the model's generalization ability in some cases. However, this can also lead to a compromised learning process, where the imputation corrupts the underlying knowledge the model is supposed to learn, ultimately resulting in artificially improved classification results that do not reflect the true nature of the data and are prone to fail in the real world.

Regarding the cardiothoracic surgery use case discussed in this chapter, while the combination of the KNNSCI method with a RF model achieved the best performance among the tested imputation methods, it did not surpass the original results obtained during the development of the risk prediction model in Section 4.3, particularly for data from 2019, as reflected in the AUROCs of Table 4.10. It is important to emphasize that the training pipeline implemented in this work differs from the one used in the original risk model development, as we aimed to apply a more standardized approach in this study. These findings underscore the value of clinician-informed imputation, which can effectively reveal underlying factors that help address critical missingness mechanisms, such as MNAR. However, when clinician-informed imputation is unavailable, the proposed correlation-based methods are a robust alternative.

4.4.6 Limitations

Despite the promising results, several limitations must be acknowledged in this study. One fundamental limitation is that while the proposed correlation-based methods effectively leverage correlations within medical datasets, it is essential to remember that correlation does not imply causation. In many cases, causal relationships between variables may provide more relevant information for imputing missing data. However, determining causality requires additional information that is often difficult to obtain. The proposed methods offer a robust alternative in scenarios where domain-specific knowledge or real-world informed imputation strategies are unavailable. Yet, future work could explore causal inference from observational data, particularly in the context of cardiothoracic surgery, to enhance the accuracy and reliability of imputation. Additionally, any imputation inherently introduces some degree of error. While unavoidable in real-world data, this imputation error could impact the downstream performance of predictive models. Therefore, it is crucial to consider the trade-off between imputation quality and the overall impact on the model's performance.

Another limitation stems from the computational cost associated with the correlation-based methods. As seen in Tables 4.12 and 4.13, the increased complexity of these techniques can lead to missing results for some datasets, highlighting the trade-off between improved imputation quality and higher computational demands. Although KNNSCI and CWRI include improvements towards computational complexity with feature selection for imputation based on correlation, further optimization may be required to make these methods more scalable for larger datasets.

Finally, while the proposed methods address the challenges of handling incomplete datasets in medical contexts, their performance was only validated on a limited set of clinical datasets. Expanding the evaluation to more diverse datasets, including those from different medical domains or healthcare systems, could provide additional insights into the generalizability of these techniques.

4.5 Providing Longitudinal Risk Predictions in Partial Annotation Scenarios¹

4.5.1 Objectives and Significance

Developing robust DSSs for cardiothoracic surgery and other medical domains often requires performing longitudinal predictions over time. In this chapter's use case, predicting the risk of complications during patient recovery extends beyond the three months we

¹This Section is based on the publication "CORKI: A Correlation-driven Imputation Method for Partial Annotation Scenarios in Multi-Label Clinical Problems" by **Ricardo Santos**, Bruno Ribeiro, Isabel Curioso, Marília Barandas, André V. Carreiro, Hugo Gamboa, Pedro Coelho, José Fragata, and Inês Sousa, published in the proceedings of the workshop *PharML: Applications of Machine Learning in Pharma and Healthcare* at the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)* (2023) [279].

have focused on thus far. The first year post-surgery remains a critical window, with high risks that significantly impact long-term outcomes [68].

Leveraging the outpatient database described in Section 4.2.2.2, new models can be developed for longitudinal prediction of complications. In the literature, time-to-event prediction models and survival analysis are often employed to estimate the likelihood or timing of an event, such as a complication, occurring within a specified period [253]. These methods typically rely on accurate time-to-event data, which track the precise timing of each event for survival or hazard function modeling.

However, in the current use case, the availability of precise time-to-event data is limited. While certain events, such as death, are accurately recorded by the nursing team at HSM, other complications may not have exact timestamps due to the periodic nature of follow-up phone calls. As a result, the longitudinal prediction of complications in cardiothoracic surgery can be reframed as a multi-label classification problem, where complications are predicted as labels associated with specific time intervals rather than being modeled based on exact event timing.

Multi-label classification is a powerful approach for addressing complex real-world problems, allowing multiple class labels to be assigned simultaneously to individual instances. This technique has been applied in various domains, such as protein function prediction in biomedical research [311], topic classification in NLP [98], and beyond. By leveraging a single learning task to predict multiple interdependent and non-mutually exclusive outcomes, multi-label classification enables the modeling of intricate relationships between class labels, which cannot be achieved with single-label classification. This is especially pertinent in clinical contexts, where different physiological manifestations are often interconnected and can benefit from being analyzed concurrently.

However, multi-label classification presents unique challenges. As the label space expands, the computational cost of training and inference tends to increase, leading to scalability issues in more complex models [111]. Additionally, multi-label datasets frequently suffer from imbalanced label distributions, where some labels are significantly more common than others [314]. In real-world data collection, annotation challenges also arise, making obtaining a comprehensive set of labels difficult or infeasible. This is particularly problematic in healthcare settings, where predicting multiple health outcomes concurrently or over different time horizons can result in incomplete or missing annotations. The cardiothoracic surgery use case suffers from such problems, where the annotation of some points in time is missing, often related to patients not answering the phone interviews. DSSs in these environments must be robust and resilient to these inconsistencies.

A common approach to addressing incomplete labels is to drop samples with missing annotations. However, this can reduce the dataset size and introduce bias, as verified in Section 4.4.5.2 with list-wise deletion. While it is noteworthy that domain knowledge can help estimate partial labels more accurately, this option is not always available. Various strategies have been proposed in the literature to overcome these challenges.

This work focuses on developing effective strategies to handle missing labels in

multi-label classification from a data-centric perspective, adaptable to various ML algorithms. With the overarching goal of enhancing robustness in the longitudinal prediction of severe complications in cardiothoracic surgery, we build on the innovations discussed in Section 4.4 to address the challenge of multi-label learning with missing labels. In this context, we introduce Correlation-Optimized and Robust K Nearest Neighbors Imputation for Multi-Label Classification (CORKI), an innovative label imputation technique that incorporates different types of feature-label correlations into a KNN strategy. CORKI is designed to strengthen the robustness of clinical DSSs by providing reliable insights that support personalized healthcare. Although CORKI applies to various domains, it has been specifically tailored to address the unique challenges of medical data, including incomplete records and highly imbalanced label distributions.

To validate the proposed method, we evaluated its performance using three public datasets related to ECG multi-label disease diagnosis, in addition to the cardiothoracic surgery dataset for long-term risk prediction. Imputation accuracy and the impact on multi-label classification tasks were assessed by comparing CORKI to conventional sample-dropping approaches and domain-knowledge imputation techniques.

4.5.2 Literature Review

The problem of multi-label learning with missing labels has been extensively studied in the literature, primarily through data-centric and model-centric methods. Data-centric approaches aim to reconstruct the label matrix before learning, while model-centric strategies integrate missing labels into the learning process using regularization techniques and adaptive loss functions [137].

From a data-centric perspective, a common approach involves leveraging inter-label relationships to infer missing labels. Huang et al. [136] utilized pairwise label correlations to impute missing labels, while Cheng and Zeng [64] incorporated both positive and negative label correlations. Recent studies have highlighted the importance of considering the feature matrix when dealing with missing labels. Rastogi and Mortaza [262] proposed a method that combines label correlations with the structural proximity of samples in the feature matrix, achieving state-of-the-art results. Zhang et al. [359] employed both feature and label embeddings to create enriched versions of label matrices without missing values. Tan et al. [312] partitioned data samples into smaller groups using granular computing principles to improve missing-label value estimation.

At the model level, Jain et al. [146] introduced a generative framework for multi-label classification that computes the likelihood of a negative value corresponding to a missing label. Similarly, Ben-Baruch et al. [30] proposed an asymmetric loss that weights contributions from imputed labels, initially performed by balancing the probability of some label in an instance and among the data. Ibrahim et al. [141] adopted a similar approach, adjusting the loss function by incorporating confidence measures to weigh the importance of missing labels. Durand et al. [81] exploited label proportions to adapt

the loss function, followed by curriculum learning strategies to predict missing labels effectively.

While these methods have been successful in various learning tasks, their applicability is often limited. Many techniques are designed for image or text data, with limited validation on other modalities such as tabular data or time series [10, 137, 146, 346, 350]. Furthermore, model-centric methods are frequently associated with complex architectures, which can hinder scalability and explainability [30, 81, 141, 146]. In healthcare scenarios, where data is often imbalanced and heterogeneous, most studies have focused solely on clinical text datasets for evaluation without addressing unique challenges such as severe label imbalance [136, 137, 359].

This work aims to bridge the identified gaps by introducing a novel data-centric method designed specifically for imbalanced clinical use cases. This method leverages a new type of correlation to enhance label imputation in multi-label learning with missing labels and incorporates a dimensionality reduction to improve scalability. CORKI extends these concepts, building on the work detailed in Section 4.4, which examined the influence of correlations between missing data patterns and known values for feature imputation [72]. The KNNSCI method, previously developed and evaluated, plays a key role as the foundation for CORKI, serving as its baseline for further enhancement and validation.

4.5.3 Methodology

In multi-label classification, each instance can be associated with multiple interdependent and non-mutually exclusive labels, making it a more complex learning task than single-label classification. Adding to the complexity of the problem, real-world multi-label classification scenarios often involve partial annotations or missing labels, a challenge that traditional algorithms are not designed to handle effectively.

External data-level processing methods, such as label imputation, have become essential in addressing this issue. Label imputation allows for the exploitation of all available data, thus ensuring that incomplete annotations do not hinder the learning process. By focusing on data-centric solutions, imputation techniques can improve the robustness of multi-label models in the presence of missing labels.

In this context, we propose CORKI, a label imputation method specifically designed for multi-label tasks with incomplete annotations. CORKI builds on the KNN imputation strategy, commonly used to replace missing values in feature matrices based on proximity measures, such as Euclidean distance [320]. However, unlike conventional methods, CORKI introduces a correlation-based dimensionality reduction step to capture only the most relevant features for the imputed label.

CORKI goes beyond the traditional correlation between variable values by incorporating correlations between values and the missingness patterns of labels. This innovative approach leverages the idea that missing labels may share underlying causes, and thus, their missingness patterns could provide valuable insights for the imputation process.

Suppose a strong correlation exists between a label’s missingness pattern and other variables’ values. In that case, this information can help narrow down the possible values for the missing elements.

Once these correlations are identified, CORKI uses them to weight the samples during the imputation process, similarly to how methods like the proposed KNNSCI or the CoHiKNN by Liu et al. [196] handle missing feature values. By integrating these correlation measures, CORKI improves the accuracy of label imputation in multi-label classification tasks. Additionally, to address the common challenge of imbalanced data in real-world clinical datasets, CORKI includes an optional weighting term that accounts for class prevalence, further enhancing its robustness in diverse real-world settings.

4.5.3.1 Datasets

The development and evaluation of CORKI were performed using two primary use cases. Alongside the longitudinal cardiothoracic surgery risk estimation, which is the focus of this chapter, a multi-label classification of cardiovascular diseases from ECG data was also employed, this one focusing on a single moment in time.

Cardiothoracic Surgery Dataset

For the longitudinal prediction of severe complications throughout the first year following cardiothoracic surgery, data from HSM was utilized. In addition to clinical variables obtained from the inpatient dataset (detailed in Section 4.2.2.1), a new label construction methodology was developed from the outpatient dataset (detailed in Section 4.2.2.2). The nursing team registered complications through phone calls at 30, 90, 180, and 365 days post-discharge. These serve as multi-label objectives, focusing on identifying patients at high risk of developing severe complications between follow-up intervals.

Missing labels primarily resulted from missed calls or the occurrence of post-mortem intervals, contributing to incomplete annotations. Table 4.14 presents key statistics of the dataset, highlighting the significant imbalance in complication prevalence (averaging 2.8%). Additionally, the number of missing samples exceeds the number of positive samples. Although the exact origin of these missing labels is unknown, they are likely influenced by either a MAR mechanism, where the missingness is associated with observed data values, or a MNAR mechanism, where the unobserved values themselves contribute to the missingness [191].

The data partition strategy follows principles similar to those in the previous sections. Given the extensive collection period, which may have seen changes in medical procedures, the external test set, including all surgeries performed in 2019, was maintained. Additionally, a standard test set was obtained by randomly selecting a patient-stratified group comprising 30% of the data from 2008 to 2018. The remaining training set was divided into five folds to run experiments with 5-fold CV.

Table 4.14: Records, label, and missing values distribution in the multi-label cardiothoracic surgery dataset.

Prediction Interval	Cardiothoracic Surgery Dataset
Total samples	5,890
Total complications (positive labels)	
[0, 30] days	196
]30, 90] days	184
]90, 180] days	141
]180, 365] days	146
Total positive labels	667
Average positive labels	166.8
Average positive class prevalence	2.8%
Total missing labels	735
Average missing labels	183.8
Average missing rate	3.1%

ECG Datasets

Regarding the multi-label classification of cardiovascular diseases, we resorted to ECG data from three distinct datasets. ECGs are non-invasive tests that record the heart's electrical activity and are widely used to diagnose cardiovascular diseases through ML-based multi-label classification techniques [221].

To evaluate CORKI, we leveraged three standardized datasets from the PhysioNet Computing in Cardiology Challenge 2020, a well-known benchmark for automated ECG classification pipelines [13]. The datasets used were the China Physiological Signal Challenge 2018 (CPSC) [193], the Physikalisch Technische Bundesanstalt XL (PTB-XL) [331], and the Georgia 12-lead ECG Challenge (G12EC) [112]. All three datasets provide 12-lead ECG signals along with demographic information (age and sex) and multi-label annotations.

Nine standard cardiovascular disease labels were selected across the datasets, and a consistent preprocessing framework was applied. The ECG signals were resampled to 250 Hz, and a 10-second window was extracted from the center of each recording to avoid the low-quality signals often present at the extremities. The signals underwent a second-order band-pass Butterworth filter (1-40 Hz) and were z-normalized.

Feature extraction was performed on the processed signals, which involved computing morphological features, wavelet descriptors, and time intervals based on R-peak detection. This resulted in 80 features per signal [222]. Additionally, demographic data such as age and sex were included. A 5-fold split was applied to each dataset using the provided folds (when available) or by implementing a random stratification approach, ensuring balance across positive class labels, age, and sex. The final fold was reserved for testing, while the

4.5. PROVIDING LONGITUDINAL RISK PREDICTIONS IN PARTIAL ANNOTATION SCENARIOS

remaining folds were used for CV. Table 4.15 details the target diseases and the available sample counts. The "normal" class (NORM) was treated as a separate label rather than as the absence of all other conditions.

Table 4.15: Records and label distribution in the CPSC, PTB-XL and G12EC datasets.

Disease	Acronym	CPSC	PTB-XL	G12EC	Total
Total samples		6,871	13,932	4,301	25,104
Total diagnoses (positive labels)					
Normal ECG	NORM	918	9,514	1,735	12,167
Atrial fibrillation	AF	1,220	1,514	568	3,302
First-degree atrioventricular block	I-AVB	722	795	766	2,283
Left bundle branch block	LBBB	235	536	231	1,002
Premature atrial contraction	PAC	614	398	636	1,648
Premature ventricular contraction	PVC	699	1,153	41	1,893
Right bundle branch block	RBBB	1,857	542	554	2,953
ST-segment depression	STD	868	1,009	38	1,915
ST-segment elevated	STE	220	28	134	382
Total positive labels		7,353	15,489	4,703	27,544
Average positive labels		817	1,721	522.4	3,060.4
Average positive class prevalence		11.9%	12.4%	12.1%	12.2%

Since these datasets contain complete annotations for all diseases, artificial missing labels were introduced to simulate real-world scenarios. To evaluate the robustness of CORKI under varying levels of incomplete labels, we randomly removed 3%, 5%, 10%, 25%, and 50% of the labels. This created a missingness pattern typical of a MCAR mechanism, where the missingness is unrelated to the data values, either observed or missing [191].

4.5.3.2 Notation

To describe the implementation of CORKI, it is essential to define some notation following a similar scheme of Section 4.4.3.2.

Consider a multi-label tabular dataset with G features, L labels, and N samples. $l_{\text{miss}} \leq L$ denotes the number of features with partial annotations. g_i represents the values of the i th feature, where $i \in \{1, 2, \dots, G\}$, and l_j the values of the j th label, where $j \in \{1, 2, \dots, L\}$. $l_{\text{miss},j}$ denotes the missingness pattern of the j th label, a binary array indicating whether the label was annotated or not for each data sample, assuming values of 0 or 1, respectively.

C_{vv} is a $(G \times L)$ matrix, which contains the correlation coefficient between every feature, g_i , and label, l_j . Considering that each label l_j may have a variable number of missing annotations, C_{vv} is computed through pairwise deletion, where the correlation between the values of each feature and the available values within each label is obtained iteratively.

This procedure removes any dependency on a complete data subset during training, both in terms of features and labels.

Analogous to C_{vv} in structure, we define C_{vm} , which contains the correlation between each feature g_i and the missingness pattern of the j th label, $l_{miss,j}$. Figure 4.8 offers a graphical illustration of these two types of correlation.

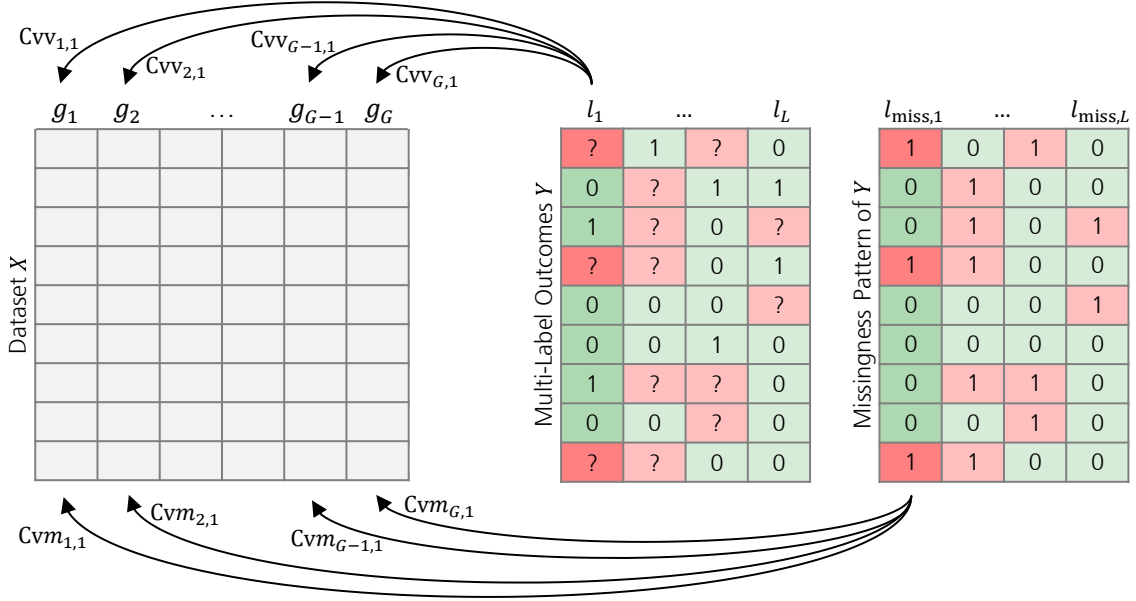


Figure 4.8: Depiction of the two types of correlation computed in CORKI. This example only looks at the correlation values for the first label, i.e., $j = 1$.

4.5.3.3 Correlation-Optimized and Robust K Nearest Neighbors Imputation for Multi-Label Classification

The CORKI method enhances traditional KNN label imputation approaches by incorporating correlation measurements between feature values and both label values and missingness patterns. By leveraging this correlation information, CORKI refines the selection of proximal samples for label estimation, ensuring that only the most relevant features are considered for distance computation. This preliminary feature selection based on correlation helps reduce data dimensionality, improving the complexity of the method for large datasets.

To address the challenge of label imbalance, common in healthcare datasets, CORKI integrates a weighting term to account for the prevalence of positive labels. This prevents redundant estimations caused by over-representing the negative class among neighboring samples. As a result, the method balances the influence of each neighbor's label, ensuring that minority classes are appropriately considered during imputation.

The following outlines the step-by-step implementation of the CORKI method:

4.5. PROVIDING LONGITUDINAL RISK PREDICTIONS IN PARTIAL ANNOTATION SCENARIOS

1. Consider a dataset X , with G features and N instances, which relates to multi-label outcomes Y , with L labels with partial annotations.
2. Compute C_{vv} , with the correlations between feature and label values, and C_{vm} , with the correlations between feature values and label missingness patterns. CORKI considers the absolute value of the correlations, thereby considering the strength of the association rather than its direction.
3. For each label l_j , select from X the subset X_{miss_j} , which includes all samples where l_j is missing, i.e., where the missingness pattern, $l_{\text{miss},j}$, assumes the value 1. Additionally, create the subset X_{obs_j} , with samples where l_j is observed, i.e., when $l_{\text{miss},j}$ is equal to 0. If X_{miss_j} is empty, skip to the next label.
4. Compute the coefficient $r_{Cj,i}$ between label l_j and the i th attribute of X using the correlation matrices. This coefficient is identical for all samples within the subset X_{miss_j} and is calculated as follows:

$$r_{Cj,i} = \frac{1}{2} \times (C_{vv}[i, j] + C_{vm}[i, j]), \quad 0 \leq r_{Cj,i} \leq 1 \quad (4.6)$$

5. Remove from X_{miss_j} and X_{obs_j} all features with $r_{Cj,i} < \text{threshold}_c$. threshold_c defines high correlation and was set as 0.7 based on Schober et al. [284]. If all features are removed, then this method cannot be applied. Still, in any case, it would not be beneficial to use a correlation-based method on a set of variables without significant correlations.
6. For each sample, find the K nearest neighbors of each sample in X_{miss_j} , within the updated subset X_{obs_j} , using the Euclidean distance.
7. For each instance of X_{miss_j} , replace the missing label l_j by a weighted mode between the K nearest neighbors' labels. The weight of the k th neighbor, $w_{j,k}$, is given by the inverse of the Euclidean distance, $d_{j,k}$, and the prevalence of its class, p_k , defined by:

$$w_{j,k} = \frac{1}{d_{j,k} \times p_k}, \quad (4.7)$$

For each label l_j , the prevalence p_k is a ratio between the instances from the class of k and the total number of cases. The distance $d_{j,k}$ was previously scaled into a range from 0 to 1, where one corresponds to the distance of the farthest neighbor among K . While the prevalence of each neighbor's class addresses dataset imbalance, it can be ignored if considered irrelevant. Then, the weight of the k th neighbor, $w_{j,k}$, only depends on the normalized Euclidean distance:

$$w_{j,k} = \frac{1}{d_{j,k}}, \quad (4.8)$$

In both cases, the weights are normalized, i.e., sum up to 1.

8. Repeat Steps 4-8 until all missing values from all labels have been imputed.

4.5.4 Experiments

To evaluate the performance and robustness of CORKI, we conducted experiments on both the multi-label cardiovascular disease classification from ECG datasets and the longitudinal prediction of severe complications in cardiothoracic surgery. These experiments assessed both the quality of label imputations and the impact of those imputations on multi-label classification performance. The CORKI method was compared to a sample-dropping approach and a domain-knowledge imputation strategy based on clinicians' inputs specific to the cardiothoracic surgery dataset.

4.5.4.1 Imputation Quality

The first experiment evaluated CORKI's imputation accuracy. This was measured by comparing the imputed values generated by CORKI against the ground truth labels in fully annotated datasets. The three ECG datasets (CPSC, PTB-XL, and G12EC) provided a basis for injecting artificial missing labels at five different rates (3%, 5%, 10%, 25%, and 50%) to simulate real-world partial annotations.

Given the binary nature of the label classes, imputation quality was computed using balanced accuracy. This methodology is distinct from Section 4.4.4.1, where the focus was also on continuous value imputation. The evaluation focused on test sets only, with the feature-label correlations derived from the training data to avoid data leakage. The cardiothoracic surgery dataset was excluded from this experiment due to the naturally missing annotations lacking a ground truth reference.

CORKI was tested with different parameter configurations, particularly regarding the number of nearest neighbors K , set at values of 5, 20, and 50. Additionally, to assess the impact of class imbalance, the experiments were conducted with and without the weighting term, as described in Section 4.5.3.3.

4.5.4.2 Classification Performance

The second experiment examined the influence of CORKI on multi-label classification performance. Four multi-label classifiers were used: DT, RF, MLP, and KNN. This experiment was conducted on both the ECG and cardiothoracic surgery datasets to compare how CORKI's imputations affected the classification tasks relative to the sample-dropping approach and, for the cardiothoracic surgery dataset, a domain-knowledge imputation rule. While the missing labels of post-mortem periods were replaced with 1, the remaining missing values assumed the value 0. This presumes similarities between deceased patients and those with complications, while unconfirmed deaths are considered low-risk periods.

We applied a comprehensive optimization pipeline to minimize external factors influencing label imputation. This included hyperparameter tuning through random search,

forward sequential feature selection, and decision threshold optimization. The optimization was performed on the complete versions of the ECG datasets and applied across all experiments for consistency. The optimization was conducted separately for each experiment for the cardiothoracic surgery dataset, where no complete data reference was available. The tested hyperparameters are available in Table A.6.

Performance was evaluated using the macro average of the AUROC. Additional metrics, including sensitivity, specificity, and F1-score, were used to assess the classification performance thoroughly. The test sets used for performance evaluation were not subject to label imputation, and in the cardiothoracic surgery dataset, initially missing labels were excluded from the metrics computation.

By comparing imputation accuracy and classification performance, these experiments comprehensively evaluate CORKI and its applicability in handling missing labels in clinical and biomedical contexts.

4.5.5 Results

4.5.5.1 Imputation Accuracy

The effectiveness of CORKI in label imputation was evaluated by measuring imputation accuracy on the test sets. Given that the labels in the ECG multi-label classification use case are binary, accuracy was calculated as the average hit-or-miss rate across all labels, weighted by each label's prevalence. As outlined in Section 4.5.3.3, two variations of CORKI were tested: the standard version, which relies solely on distance, and a prevalence-weighted version that accounts for label imbalance during imputation.

The results, presented in Table 4.16, include various missing rates and datasets and are reported as the mean and standard deviation across all labels (i.e., diseases) for different K values of neighbors used by CORKI. The "Total" column represents the overall imputation accuracy performance, computed as the average and pooled standard deviation across all datasets.

Across all datasets, the highest average imputation accuracy of $73.9\% \pm 15.1\%$ was achieved by the prevalence-weighted CORKI at a missing rate of 5%. This result highlights the benefit of incorporating class prevalence into the imputation process, particularly in scenarios with moderate levels of missing labels. Conversely, the lowest accuracy, $58.2\% \pm 10.1\%$, was observed for the non-weighted CORKI at a missing rate of 50%, indicating that larger proportions of missing labels significantly degrade performance.

Among the datasets, PTB-XL consistently yielded the best imputation results across all missing rates, achieving a peak accuracy of $78.9\% \pm 13.4\%$ at a missing rate of 3%. This superior performance is likely attributable to the larger sample size of PTB-XL compared to the other datasets (see Table 4.15), which enhances the reliability of correlation and proximity measures utilized by CORKI.

The datasets exhibit significant class imbalance, with an average positive class prevalence of 12.2%. The prevalence-weighted version of CORKI outperformed the standard

Table 4.16: Results of ECG imputation for the test set in the CPSC, PTB-XL, and G12EC datasets. The Total column presents the average and pooled standard deviation of imputation accuracy across all datasets. Results are presented as a percentage (%) in terms of balanced accuracy. CORKI P. refers to the disease prevalence weighting.

Missing Rate	Strategy	CPSC	PTB-XL	G12EC	Total
3	CORKI	67.8 $\pm$ 20.6	63.7 $\pm$ 17.3	62.8 $\pm$ 17.9	64.9 $\pm$ 18.6
	CORKI P.	71.7 $\pm$ 11.6	78.9 $\pm$ 13.4	68.8 $\pm$ 24.0	73.1 $\pm$ 17.2
5	CORKI	61.5 $\pm$ 15.6	64.6 $\pm$ 17.5	60.6 $\pm$ 13.3	62.2 $\pm$ 15.6
	CORKI P.	73.3 $\pm$ 10.3	78.1 $\pm$ 12.6	70.4 $\pm$ 20.4	73.9 $\pm$ 15.1
10	CORKI	58.1 $\pm$ 9.2	64.3 $\pm$ 17.1	59.4 $\pm$ 12.9	60.6 $\pm$ 13.5
	CORKI P.	70.6 $\pm$ 12.3	78.6 $\pm$ 12.6	67.2 $\pm$ 16.4	72.1 $\pm$ 13.9
25	CORKI	60.3 $\pm$ 9.1	57.9 $\pm$ 11.2	58.8 $\pm$ 12.0	59.0 $\pm$ 10.8
	CORKI P.	72.6 $\pm$ 8.8	70.9 $\pm$ 14.4	67.2 $\pm$ 11.0	70.2 $\pm$ 11.6
50	CORKI	58.5 $\pm$ 8.8	57.2 $\pm$ 10.7	58.8 $\pm$ 10.8	58.2 $\pm$ 10.1
	CORKI P.	68.9 $\pm$ 9.1	69.3 $\pm$ 13.3	68.4 $\pm$ 9.74	68.9 $\pm$ 10.9

version regarding balanced accuracy, demonstrating greater resilience to imbalanced data. However, the similar magnitude of standard deviations indicates variability in performance across different labels, which suggests that some labels are inherently more challenging to impute accurately.

As anticipated, imputation accuracy declines as the missing rate increases due to the diminishing availability of information needed to establish meaningful correlations. Interestingly, standard deviations tend to decrease with a higher missing rate, indicating a more consistent imputation performance across labels despite the drop in overall accuracy. The performance drop for the prevalence-weighted CORKI is relatively modest, with a decline of only 5.0 percentage points between the 3% and 50% missing rates.

4.5.5.2 Classification Performance

While label imputation strategies can mitigate the adverse effects of partial annotations, evaluating their impact on classification performance is essential, as label distributions may shift, leading to varied outcomes. We assessed the performance of ECG classification and longitudinal risk prediction in cardiothoracic surgery under different imputation conditions. The public datasets used for ECG classification had synthetically injected missing labels, enabling the analysis of various missing rates, while the cardiothoracic surgery dataset featured real-world missing labels at an average natural rate of 3.1% (see Table 4.14).

Table 4.17 summarizes the classification performance of the ECG use case across different missing rates. Each set of results corresponds to the pipeline that achieved the

4.5. PROVIDING LONGITUDINAL RISK PREDICTIONS IN PARTIAL ANNOTATION SCENARIOS

highest CV AUROC and the best number of neighbors K . For each dataset, the prevalence-weighted and unweighted versions of CORKI were compared against the original data and the sample-dropping approach. In addition to the AUROC, other performance metrics were computed, including sensitivity, specificity, and F1-score, following threshold optimization. These extended results are available in Table A.7.

Table 4.17: Results of ECG classification for the test set in the CPSC, PTB-XL, and G12EC datasets. Results are presented as percentage (%) in terms of AUROC, as the mean and standard deviation of all labels. CORKI P. refers to the disease prevalence weighting and K denotes the number of neighbors utilized by CORKI.

Missing Rate	Strategy	K	CPSC	PTB-XL	G12EC
3	Original	-	92.9 ± 3.9	91.6 ± 8.6	86.5 ± 13.4
	Drop	-	92.9 ± 3.9	90.6 ± 11.1	85.7 ± 12.1
	CORKI	50	93.0 ± 3.9	90.3 ± 12.4	86.2 ± 13.9
	CORKI P.	5	92.5 ± 4.0	90.3 ± 11.5	85.4 ± 14.8
5	Drop	-	92.6 ± 4.1	90.7 ± 10.1	85.1 ± 13.8
	CORKI	50	93.0 ± 3.8	90.5 ± 12.2	85.9 ± 14.6
	CORKI P.	5	92.4 ± 4.0	90.3 ± 12.1	84.8 ± 15.9
10	Drop	-	92.0 ± 4.0	85.8 ± 16.2	82.4 ± 16.1
	CORKI	20	92.7 ± 3.9	90.8 ± 10.8	86.2 ± 13.7
	CORKI P.	5	92.3 ± 3.8	90.3 ± 11.2	84.4 ± 15.0
25	Drop	-	$81.0 \pm 9.2^*$	-	65.7 ± 22.9
	CORKI	50	92.1 ± 4.4	90.2 ± 11.9	84.8 ± 16.5
	CORKI P.	50	89.4 ± 5.2	89.8 ± 10.2	82.7 ± 13.2
50	Drop	-	-	-	-
	CORKI	50	$87.4 \pm 6.6^*$	90.0 ± 10.1	83.7 ± 15.9
	CORKI P.	5	87.4 ± 5.9	87.7 ± 10.9	81.4 ± 14.0

* KNN was the best-performing model. MLP was selected otherwise.

Generally, the complete dataset provided the best classification performance, with AUROC scores surpassed by no more than 0.1%. In some cases, imputation strategies produced better results than those achieved with the original dataset, as artificial values may better align with expected data distributions [72]. Among the classifiers tested (DT, RF, MLP, and KNN), MLP consistently outperformed the others, except in two scenarios. However, the optimal K value did not follow a clear trend, though using five neighbors was the most common.

CORKI delivered performance comparable to the complete dataset in most cases up to a missing rate of 25%, while the sample-dropping approach maintained such performance only up to 10% missing rate. As instances with missing labels were removed, the resulting data distribution became increasingly biased, leading to sub-optimal learning

outcomes. This effect was more pronounced at higher missing rates, where sample dropping significantly degraded performance. At a missing rate of 50%, sample dropping eliminated all training samples in some cases, leaving CORKI as the only viable option.

Interestingly, contrary to the first experiment, the prevalence-weighted version of CORKI did not outperform the standard version in classification performance. In fact, the weighting decreased specificity, particularly in the CPSC and PTB-XL datasets, as it tended to increase sensitivity at the expense of false positives. This resulted in a lower F1-score, as more positive labels were imputed, skewing the balance between true positives and false positives.

Table 4.18 presents the classification results for the longitudinal risk prediction use case using the cardiothoracic surgery dataset in terms of the AUROC. Additional metrics are aggregated in Table A.8. Two test sets were considered: a random split (2008-2018) and a set comprising all surgeries from the most recent year available (2019). The dataset’s severe imbalance, with a natural missing rate of 3.1%, posed additional challenges. MLP models with K set to 50 consistently achieved the highest AUROC scores, although F1-scores remained low due to the tendency to favor the positive class, which led to an increase in false positives.

Table 4.18: Classification results of longitudinal risk estimation in the cardiothoracic surgery test sets. Results are presented as percentage (%) in terms of AUROC as all labels’ mean and standard deviation. CORKI P. refers to the disease prevalence weighting. MLP was always selected as the best model. * corresponds to the same outcome in all three tested K .

Strategy	K	Cardiothoracic Surgery	
		2008-2018	2019
Drop	-	57.3 $\pm$ 2.6	55.6 $\pm$ 3.5
Domain	-	67.9 $\pm$ 4.7	60.5 $\pm$ 3.2
CORKI	*	66.3 $\pm$ 4.1	58.6 $\pm$ 1.6
CORKI P.	50	67.9 $\pm$ 5.9	57.0 $\pm$ 3.7

In addition to sample dropping, a domain knowledge-based imputation strategy was tested, which yielded the best AUROC scores. While sample dropping resulted in the poorest performance, the domain knowledge approach produced the most accurate predictions. Notably, in the first test set, the prevalence-weighted version of CORKI matched the domain knowledge approach in AUROC, albeit with higher sensitivity. CORKI consistently outperformed the sample-dropping method, particularly in the first test set, a less evident trend in the public datasets at the same missing rate of 3%. This discrepancy likely stems from the extreme imbalance of the cardiothoracic surgery dataset, where positive samples account for only 2.8% of the data, making the loss of any positive samples through sample dropping disproportionately impactful.

All methods exhibited reduced performance when evaluated on the 2019 test set,

indicating the presence of temporal drift in either the features or labels. Interestingly, the prevalence-weighted version of CORKI performed worse than the standard version, suggesting that the class distribution may have shifted over time, diminishing the benefits of applying prevalence weighting and likely affecting the quality of imputations.

4.5.6 Discussion

Handling partial annotations in multi-label classification poses significant challenges, especially in real-world scenarios where missing labels often arise from human decisions, machine errors, or inherent limitations of data collection processes, such as crowdsourcing annotations. CORKI addresses these challenges by introducing a label imputation method that leverages feature-label correlations, distance metrics, and class prevalence to improve label imputation accuracy and enhance multi-label classification performance.

Our evaluation of CORKI across the ECG multi-label use case demonstrated that the method performs well in imputation accuracy, particularly in high class imbalance. Notably, the weighted version of CORKI consistently yielded more accurate imputations at varying missing rates, highlighting the importance of accounting for class prevalence when addressing partial annotations. However, the performance evaluation was limited to three datasets with similar characteristics, and further testing on a broader range of datasets is necessary to validate the generalizability of these findings.

Regarding classification performance, CORKI exhibited robust results, comparable to those achieved with the original datasets, despite the inherent loss of information due to missing labels. The method's advantages became more pronounced as the missing rates increased, where standard sample-dropping approaches struggled. However, the prevalence-weighted version of CORKI produced mixed results, with performance varying across datasets and different K neighbors. This variability suggests that prevalence weighting and K should be carefully optimized for each specific application. Future research could explore the impact of these parameters in greater depth, examining their effects on different types of multi-label classification problems and their interactions with data characteristics such as imbalance and missing rate.

Another important observation is the relatively low F1-scores, which can be attributed to the severe class imbalance in the datasets. While AUROC was chosen as the primary metric due to its balance between sensitivity and specificity, the low F1-scores indicate that improvements in precision may be necessary for specific use cases. Addressing this in future studies, particularly in healthcare contexts where precision can be critical, would be beneficial.

From a healthcare perspective, and especially focusing on this chapter's cardiothoracic surgery use case, CORKI's impact is particularly relevant. Real-world healthcare datasets often feature intermittent data collection and missing outcomes due to patient monitoring and follow-up complexities, and human-informed correction may not always be available. Traditional multi-label classification pipelines struggle to handle such scenarios,

making accurate label imputation methods crucial. CORKI’s approach, which leverages relationships between variables and missingness patterns, offers a more context-aware method for imputing missing labels. This is especially valuable in healthcare settings where sensitivity often takes precedence over specificity, as the cost of false negatives (e.g., missing a critical complication) can be much higher than false positives. Nonetheless, further work is needed to ensure that CORKI’s predictions are trustworthy, potentially by incorporating uncertainty estimation techniques.

4.5.7 Limitations

Despite CORKI’s promising results, several limitations must be acknowledged. First, the current evaluation was limited to a small number of datasets with similar characteristics, focusing mainly on ECG data and longitudinal risk prediction in cardiothoracic surgery. Future work should involve a more comprehensive evaluation across a broader range of datasets, including different domains and data modalities, to validate the robustness and scalability of the method. Testing against existing methodologies in the literature, such as the work by Rastogi and Mortaza [262] and Ben-Baruch et al. [30], could provide further insights into CORKI’s comparative performance. The number of ML models tested was not extensive either, but the algorithms adapted to multi-label classification scenarios are limited.

Another key limitation is the lack of exploration of model-based strategies for handling missing labels, which have been proposed in other studies. While CORKI focuses on a data-centric approach, incorporating model-level adaptations, such as those involving specialized loss functions and regularization techniques, could enhance the method’s effectiveness. Model-centric approaches will be introduced and discussed further in this thesis, specifically in Chapter 6.

Furthermore, while CORKI’s reliance on correlation-based imputation has shown promising results, it is crucial to recognize that correlation does not imply causation. Although causal relationships may offer more meaningful insights, they are often difficult to identify due to the lack of comprehensive domain knowledge or access to detailed data. In healthcare, where observational data is prevalent, future research could explore causal inference methods to complement the correlation-based techniques introduced by CORKI, similar to the future work in Section 4.4.6. This would help improve imputation accuracy and the reliability of DSSs in clinical settings.

Lastly, all imputation methods inevitably introduce some error, which can accumulate and affect overall model performance. While imputation is often unavoidable in real-world data, future work should aim to quantify and mitigate the impact of imputation-related errors, particularly in critical applications such as healthcare.

In summary, CORKI presents a novel and promising approach to handling missing labels in multi-label classification, particularly in clinical contexts. However, addressing its limitations through broader testing, exploring model-based strategies, and integrating

causal inference methods will be essential in advancing this work and enhancing the robustness of DSSs in healthcare and beyond.

4.6 Integration and Discussion

This chapter has investigated the complexities of developing robust risk models using real-world data, focusing on the cardiothoracic surgery use case in collaboration with clinicians and nurses from HSM under the CardioFollow.AI project. Our work has encompassed the entire pipeline, from objective formulation and data preparation to model construction, deployment, and real-world evaluation.

The core of our contribution lies in the design, development, and implementation of an AI-based DSS aimed at improving patient outcomes following cardiothoracic surgery. Unlike much research that remains theoretical or is restricted to processed open datasets, our work has embraced a real-world application. By closely partnering with clinicians and nurses, we ensured that domain knowledge was seamlessly integrated into the development of the system. This collaboration allowed us to create a robust DSS for predicting complications within three months post-surgery, which was then used to prioritize patients for remote monitoring. The involvement of clinical experts undoubtedly improved the system's accuracy and reliability. However, we also recognized that expert participation is not always feasible or efficient, especially in larger-scale or resource-limited environments.

Acknowledging the common challenge of missing data in real-world datasets, we introduced innovative methods to incorporate correlation into reliable missing value imputation. We developed and validated three novel approaches, CWKNNI, KNNSCI, and CWRI, within the cardiothoracic surgery use case. These methods capitalize on the inherent relationships between variables to enhance the robustness of imputation and, consequently, the performance of predictive models. Our focus then extended to the issue of missing labels, often arising from incomplete outcome registration. In addressing this problem, we proposed CORKI, a method for missing label imputation in multi-label classification tasks. CORKI was validated within the same cardiothoracic surgery context, targeting the prediction of severe complications in the year following surgery.

Additionally, this chapter introduced a new challenge in the multi-label classification domain: predicting cardiovascular diseases from ECG data. This will be explored in greater depth in the following chapter, broadening the scope of our work and further demonstrating the applicability of our methodologies to different healthcare contexts.

Throughout this journey, we have contributed to the scientific community by proposing new techniques and insights into building more robust DSSs that can be deployed in real-world settings and ultimately improve health outcomes, particularly for an increasingly aging population. Our approach has been human-, data- and deployment-centric, emphasizing the importance of robust, trustworthy systems grounded in real-world needs and constraints.

Nevertheless, significant challenges remain. While the results achieved were satisfactory for the specific needs of our use case, there is still room for improvement in the robustness of complication prediction. Although we extensively explored traditional ML methods and incorporated valuable clinician feedback, new strategies such as DL or attention-based models may offer further advancements. Additional data, such as multi-modal sources (e.g., clinical registries, imaging, or electrophysiological exams), could also unlock hidden predictive power. Furthermore, including historical patient information may provide insight into critical events that could influence future outcomes. Another crucial consideration is the sustainability of these models over time. Performance degradation due to temporal data shifts is a significant challenge, and manual periodic retraining may not be a viable long-term solution.

Future work should build upon these contributions and further refine the requirements for robust DSSs in healthcare. Engaging domain experts is vital to ensure that models are aligned with clinical practice and patient needs. Although the HSM use case offers a rich and exemplary framework, it represents a single hospital, and broader generalizability will require validation across multiple institutions and larger cohorts. Additionally, integrating uncertainty into decision-making processes and enhancing explainability will be critical for increasing trust in AI-based clinical DSSs.

4.7 Final Remarks

This chapter has addressed one of the central objectives of this thesis: the development of robust DSSs in real-world medical contexts, particularly focusing on cardiothoracic surgery. Through data-, human-, and deployment-centric approaches, we tackled the unique challenges posed by incomplete and heterogeneous datasets and the difficulty of predicting post-surgical complications. Our work has shown that leveraging domain knowledge, where feasible, significantly enhances the quality of predictive models. However, we also recognized that relying solely on domain information is not always practical or even possible, particularly when scaling solutions across different settings.

The challenges presented by this problem motivated us to explore different use cases where robustness can be tested and enhanced in continuous and multimodal contexts. The nature of the available data, constrained by real-world collection practices and incomplete annotations, highlighted the limitations of current methodologies and the need to investigate alternative approaches. This experience inspired us to develop new strategies for dealing with data issues such as missing labels and missing inputs. The summary of contributions from Chapter 4 is shown in Figure 4.9.

The contributions developed in the following chapters extend the work by exploring new use cases and methodologies. Chapter 5 will work on ECG data and multimodal data fusion techniques, while Chapter 6 focuses on robustness in continuous patient monitoring, particularly handling asynchronous inputs. These advancements will address the limitations of traditional models on available data and generalization capabilities across

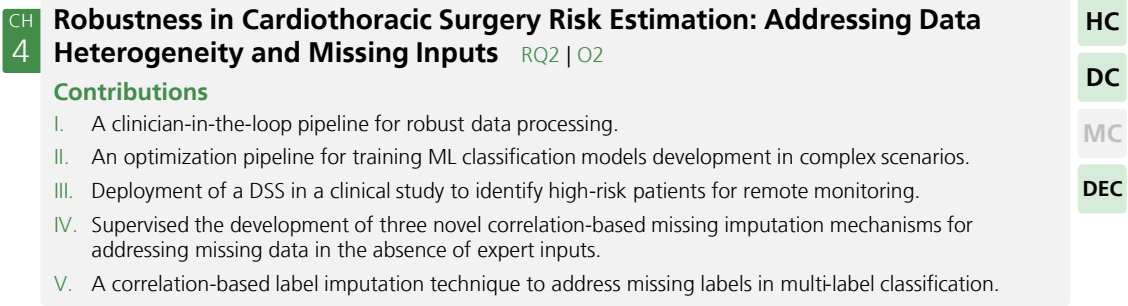


Figure 4.9: Overview of Chapter 4 contributions, aligned with the respective approaches of the holistic framework proposed in Chapter 3.

different medical domains. New use cases and methods will make predictive models more resilient in the face of real-world data challenges.

In summary, while domain-specific knowledge remains invaluable, this chapter has demonstrated that robust DSSs can be developed even in its absence by leveraging data-centric techniques. As we move forward, the learnings from this chapter will inform our exploration of continuous and multimodal data, bringing us closer to the overarching goal of building adaptable, reliable, and effective AI models.

ROBUSTNESS IN ECG DIAGNOSIS: EVALUATING MULTIMODAL DATA FUSION TECHNIQUES

5.1 Introduction

Medical data is inherently multimodal, derived from various data types, physiological phenomena, and distinct sources. This diversity presents opportunities and challenges, particularly in building robust AI models for biomedical applications. Robustness is critical for AI-driven DSSs in healthcare, where integrating and processing multiple data modalities can significantly enhance diagnostic accuracy and patient outcomes [5].

Traditional ML requires a structured input matrix of independent variables, which is not achievable in images or signals without feature extraction processes. Libraries such as the TSFEL have been proposed to extract features from multivariate signals automatically [25], which can outperform DL-based feature extraction in some use cases [34]. However, this simplification can compromise the robustness of the models, particularly in scenarios requiring rich and varied data inputs. The cardiothoracic surgery use case discussed in Chapter 4 exemplifies this challenge. While we made concerted efforts to incorporate available data sources, we could not include additional modalities, such as clinical textual notes. This exclusion likely impacted the overall performance of the DSS, highlighting a significant margin for improvement in the model's capacity to estimate risks.

Cardiovascular disease diagnosis presents another critical challenge. As one of the leading causes of death worldwide, with 20.5 million deaths in 2021, [189], one-third of the total, effective diagnosis and treatment are crucial. ECGs are widely used as a non-invasive method to diagnose cardiovascular conditions by recording the heart's electrical activity. Traditional ML methods for interpreting ECGs have relied on techniques like morphological feature extraction, but DL models now offer the potential to directly analyze raw ECG time series data from different leads, providing a more nuanced understanding of the signals [347].

Improving the robustness of AI-driven DSSs goes beyond single-modality data processing. By incorporating multimodal information, such as different interpretations of ECG signals, using both original time series and signal-to-image transformations [9], AI models could become more accurate and reliable. This chapter presents our work into multimodal AI, particularly DL models, and explores the challenges and potential of multimodal data fusion. We will investigate how these concepts can be applied to develop more robust AI models for the multi-label classification of cardiovascular diseases, specifically evaluating whether combining multiple representations of ECG data in a model-centric approach can improve model performance.

5.2 Fundamentals of Multimodality

In healthcare, data often comes from multiple sources and reflects diverse physiological phenomena, such as clinical notes, imaging, sensor readings, and lab tests (see Section 3.2). Developing AI systems capable of processing and integrating such heterogeneous data presents numerous challenges but also offers significant opportunities for enhancing clinical DSSs. For instance, an ideal robust AI model for diagnosing cardiovascular diseases may need to combine information from ECG signals, medical images, and patient history to provide accurate predictions.

By leveraging multimodal data, AI systems can achieve a deeper understanding of patient conditions and enhance diagnostic accuracy. However, integrating data from different modalities requires overcoming technical challenges related to data representation, fusion, and alignment, among others. This section will explore the concepts and challenges of multimodal AI.

5.2.1 Definitions

A *modality* refers to a specific way in which a phenomenon is perceived, experienced, or expressed, often through sensory inputs such as vision, sound, or touch [23, 188]. In the context of AI, a modality represents a distinct source of information captured through sensors or devices, such as speech and audio through microphones, images via cameras, or electrophysiological signals through sensors. Abstractions from raw modalities compose new modalities, such as language extracted from speech recordings or objects retrieved from an image [188]. For instance, in healthcare, modalities might include physiological signals like heart rate or respiratory data and visual inputs like medical imaging and their segmented organs and structures.

In AI research, the term *multimodal* describes tasks or datasets that involve multiple such modalities [23]. These modalities are typically heterogeneous, meaning they vary in structure, quality, and representation. Following the definition by Liang et al. [188], multimodal refers to the heterogeneous and interconnected nature of these multiple data

sources, where each modality contributes complementary information that interacts to enhance the overall understanding of a task.

Multimodal systems must address complex connections and interactions between modalities [188]. Modality connections refer to shared information across different modalities. For example, text and images might convey the same underlying meaning in a task like image captioning but through different representations. Modality interactions, on the other hand, emerge when combining modalities as they work together to generate new insights. For instance, when vision and language are integrated for an AI-based DSS that describes and evaluates a medical image, the system connects the data and exploits interactions to derive a richer understanding of the task at hand.

Thus, studying multimodal AI involves developing models that can handle heterogeneity, find connections, and facilitate interactions.

5.2.2 Technical Challenges in Multimodal AI

In multimodal AI, integrating and processing information from different data sources presents significant challenges in several domains. Over the years, researchers have proposed various taxonomies to categorize these challenges, aiming to understand better how to handle the intricacies of multimodal data in AI systems.

In 2019, Baltrušaitis et al. [23] laid out a foundational taxonomy that identifies five primary challenges associated with multimodal AI:

- **Representation:** This challenge includes research into effectively representing data from multiple modalities in a way that exploits their complementarity and redundancy.
- **Translation:** Refers to mapping data from one modality to another, such as generating an image description. This task is complex due to its open-ended nature, as there might be multiple valid translations, and choosing the best one is often subjective and context-dependent.
- **Alignment:** Involves identifying correspondences between elements from different modalities, such as linking spoken words to corresponding video frames, which is difficult due to issues like long-range dependencies between modalities and ambiguities in segmentation.
- **Fusion:** Refers to integrating information from multiple modalities to make predictions or decisions. Modalities contribute varying amounts of predictive power, and some may have missing or noisy data.
- **Co-learning:** Co-learning is about transferring knowledge between different modalities, necessary when one modality has limited data, and knowledge learned from another can help improve the model.

Liang et al. [188] redefined the taxonomy, introducing six challenges in light of recent advancements in multimodal AI research, such as the rise of transformer architectures and generative models:

- **Representation:** This remains the central challenge but is expanded into three sub-categories: fusion; coordination, where modalities exchange information to improve contextualization with no reduction; and fission, which encompasses creating a more extensive set of representations to reflect internal data structures, such as clusters or factors within the data.
- **Alignment:** Similar to the previous taxonomy, this challenge is about mapping connections and interactions between modality elements.
- **Reasoning:** A new addition, reasoning focuses on the ability of AI models to perform inference across modalities. This includes composing knowledge from multiple sources, understanding how multimodal data relates to intermediate concepts, and leveraging external knowledge to support reasoning.
- **Generation:** As generative models like GPT-4 [235] arose, the generation challenge involves learning processes to produce new data across modalities. This encompasses tasks such as summarization, translation, or creation.
- **Transference:** This challenge builds on Baltrušaitis' co-learning concept [23]. It involves transferring knowledge across modalities, with methods like cross-modal transfer, by applying knowledge from one modality to a task in another; co-learning, when sharing representation spaces across modalities to improve performance; and model induction, when unimodal models transfer information between them.
- **Quantification:** This challenge deals with understanding the complexities of multimodal data through empirical and theoretical analysis. This challenge involves studying how data heterogeneity influences learning and modeling and how connections and interactions between modalities are captured by AI models.

Both taxonomies address the inherent complexity of multimodal AI but reflect different stages of the field's development.

5.3 Fusion Techniques

In this thesis, we address the challenge of fusion in multimodal AI, specifically as it applies to combining and integrating different representations of ECG data. Fusion, in this context, refers to merging information from various modalities or representations to improve the overall performance of predictive models.

According to Huang et al. [138], fusion strategies in multimodal learning can be broadly categorized into early fusion, joint (or intermediate) fusion, and late fusion. Each

strategy incorporates information at different stages of the learning pipeline, and each has distinct advantages and challenges depending on the complexity and type of data involved.

Early Fusion

Early fusion, also known as feature-level fusion, combines raw data or features from multiple input modalities into a single representation before it is passed into a model for training. Different modalities are joined at the feature level in this strategy, which can be done using methods such as concatenation, pooling, or applying gated units.

Huang et al. [138] further differentiate early fusion into two types. While type I involves directly fusing original features, such as raw data from multiple ECG leads, into a combined input feature vector, type II fuses learned representations, extracted features, outputs of imaging analysis software, or neural network activations.

This strategy is often used when the data from different modalities are temporally aligned or share a similar structure, enabling a unified representation at the feature level.

Joint Fusion

Joint fusion, or intermediate fusion, integrates learned feature representations from different modalities at intermediate stages of NNs. Unlike early fusion, joint fusion propagates the loss function back to the feature extraction networks during training, ensuring the process is optimized for the final task.

Huang et al. [138] also classify joint fusion into type I, when learned feature representations are extracted from all modalities and then fused, with loss propagated back to the feature extraction layers during training, or type II, when not all input features require an explicit feature extraction step. Here, some features can be directly combined with the intermediate learned representations.

This strategy is often implemented with DL networks due to their ability to propagate loss effectively across multiple layers and modalities.

Late Fusion

Late fusion, or decision-level fusion, refers to the combination of predictions from multiple models, each trained on a different modality. Each modality contributes a separate prediction in this approach, and these predictions are aggregated using techniques such as averaging, majority voting, weighted voting, or meta-classifiers [138].

5.4 Improving Cardiovascular Disease Diagnoses using Electrocardiograms¹

*This has been joint work with
Mariana Dias (1D models) and
Hemaxi Narotamo (2D models).*

5.4.1 Objectives and Significance

Cardiovascular diseases are a leading cause of death worldwide, with their impact steadily increasing over time. In 1990, these conditions were responsible for 12.1 million deaths globally, but by 2021, this number had surged to 20.5 million [77]. While preventive measures throughout life can improve outcomes, early diagnosis remains critical for effective treatment. The ECG, a non-invasive test that records the heart’s electrical activity, is a vital tool in diagnosing various cardiovascular diseases [197]. However, manual ECG analysis is time-consuming and challenging, even for experienced clinicians [3]. Automated ECG interpretation methods can significantly ease this burden by providing faster, more accurate, and less subjective assessments, ultimately supporting more effective clinical decision-making [197].

Traditional ML approaches to ECG classification often rely on hand-crafted features, which are laborious to create and may limit the model’s ability to generalize across different patient populations [3, 218]. DL offers a powerful alternative by automatically extracting relevant features from raw data. Consequently, DL-based ECG classification has demonstrated superior performance, leading to a surge in DL methodologies tailored for ECG analysis [197]. While most DL approaches rely on 1D ECG signals, recent advancements in DL have also proposed alternative interpretations of ECG signals by transforming them into images using statistical methods [9]. These transformations enable the application of 2D methods to ECG data, offering a new perspective on the analysis of heart activity. Studies on this new perspective suggest that 2D models can sometimes outperform 1D models by capturing spatial patterns [3, 197].

Different data representations (e.g., 1D and 2D) provide complementary information, so integrating multiple modalities may enhance diagnostic accuracy. Multimodal DL leverages this concept by combining data from various sources to provide a more comprehensive analysis [23]. In this section, we evaluate the performance of combining different ECG signal representations in a multi-label cardiovascular disease classification task.

This work explores the potential of using 2D image transformations of ECG signals, specifically applying the methods proposed by Ahmad et al. [9] on the PTB-XL ECG dataset [331], which includes multiple heartbeats per record. By analyzing the performance of

¹This Section is based on the publication “Deep Learning for ECG Classification: A Comparative Study of 1D and 2D Representations and Multimodal Fusion Approaches” by Hemaxi Narotamo[†], Mariana Dias[†], Ricardo Santos[†], André V. Carreiro, Hugo Gamboa, and Margarida Silveira, published in the journal *Biomedical Signal Processing and Control* (2024) [227]. [†] First authors.

these 2D representations, we evaluate their utility in the context of cardiovascular disease diagnosis. Additionally, **our work investigates the effectiveness of multimodal fusion techniques that combine both 1D signal data and 2D image transformations**. This approach offers new insights into using hybrid models for medical signal processing and assesses whether merging different representations of ECG signals can enhance classification performance. Moreover, this study comprehensively compares various DL methodologies for cardiovascular disease classification, including state-of-the-art models with attention mechanisms. These comparisons serve as a valuable benchmark for future research and guide the selection of optimal models for diagnosing specific conditions. Figure 5.1 summarizes the compared approaches.

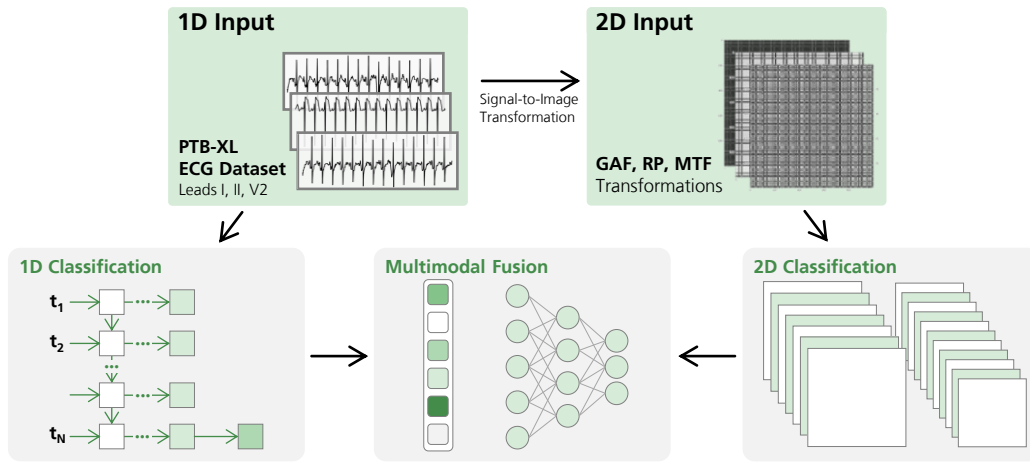


Figure 5.1: Scheme of the proposed ECG signal classification approaches. GAF, RP, and MTF represent the signal-to-image transformations.

5.4.2 Literature Review

5.4.2.1 1D Approaches for ECG Classification

Classifying ECG signals has been an intense research topic. Traditional statistical and ML models were used with features extracted from the signals. The ECG presents distinct waves called P, QRS, and T-U waves, which were identified through the manual extraction of morphological features [334]. However, the potential of DL models to avoid the need for an intense feature engineering task has shifted the focus in the literature.

RNNs are commonly used for processing sequential data, including 1D physiological signals like ECGs. RNNs use feedback connections to retain a "memory" of previous inputs, enabling the network to maintain information over time [271]. Two widely used variants of RNNs are LSTM [133] networks and GRUs [67]. LSTMs have a gated structure that allows them to handle long sequences by regulating the flow of information through input, output, and forget gates [133]. GRUs offer a more computationally efficient architecture while retaining similar functionality [66].

Both LSTM and GRU networks have been applied to ECG signal processing and analysis, demonstrating their effectiveness in tasks such as R-peak detection [176], heart disease classification [195], arrhythmia detection [352], atrial fibrillation identification [274], and heartbeat classification for continuous monitoring [273].

One-dimensional CNNs have also shown significant success in physiological signal analysis. They have effectively detected conditions like sleep apnea from photoplethysmography data [336], and segmented patterns in ECG and inertial data for pattern matching [210]. Often, 1D CNNs are combined with GRUs to enhance performance, as CNNs excel in extracting local features, while RNNs capture temporal dependencies in the data [365, 105]. This combination has been used in various applications, including ECG signal classification [118].

In the context of the PTB-XL dataset [331], several works have explored 1D-based approaches. For example, Śmigiel et al. [297] used a combination of 1D CNN and entropy-based features for ECGs classification, achieving an average accuracy and sensitivity of 76.5% and 66.2%, respectively. Strodthoff et al. [302] compared different DL models, attaining an AUROC of 93.4% when using a residual network [332]. Class imbalance is a critical challenge in ECGs classification. Virgeniya et al. [329] addressed the class imbalance problem by proposing a GRU-based model combined with an Extreme Learning Machine, improving results by applying a synthetic sampling technique. The attained sensitivity and specificity reached 91.0% and 86.8%, respectively.

Recent developments include a self-supervised method proposed by Lai et al. [175], which uses 1D ECGs signals from a specific wearable to classify 60 conditions, optimized for deployment on mobile devices. Similarly, Yang et al. [351] introduced a multi-scale network based on the Residual Neural Network (ResNet), incorporating attention modules to combine signals from different ECGs leads. Using the PTB-XL datasets, the authors obtained an overall AUROC and sensitivity of 93.25% and 81.29%, respectively.

5.4.2.2 2D Approaches for ECG Classification

CNNs have become prominent for tasks like image segmentation, recognition, and classification due to their enhanced performance [55, 299]. In ECG classification for cardiovascular disease diagnosis, researchers have begun converting 1D ECG signals into 2D images, which are then processed by CNNs for classification [9, 197, 299].

A common approach involves transforming the 1D ECG signal into a 2D time-frequency spectrogram using the Short Time Fourier Transform (STFT), enabling the integration of time and frequency details in a single image. For example, Huang et al. [135] used a 2D CNN to classify five types of arrhythmia, demonstrating superior performance over a 1D CNN, even when disregarding manual preprocessing steps, such as signal filtering and feature extraction. The final accuracy was 99.0%. Other studies have followed similar approaches. Diker et al. [80] utilized spectrogram images as input for CNNs like VGG-16, ResNet-18, and AlexNet, with AlexNet achieving the best results with an accuracy of

83.8%. Similarly, Ullah et al. [322] trained a 2D CNN to classify ECG images into eight different arrhythmia classes. However, these studies did not compare the performance of 2D CNNs against models using the original 1D ECG signal.

Ahmad et al. [9] introduced three new statistical methods, Gramian Angular Field (GAF), Recurrence Plot (RP), and Markov Transition Field (MTF), to convert ECG signals into images while preserving temporal details. Using two datasets, they transformed each 1D heartbeat into three images and achieved superior classification performance compared to time-frequency-based methods like STFT.

5.4.2.3 Multimodal Fusion

Research in cardiovascular disease classification using multimodal representation of the ECG signals is more scarce. Golrizkhatami and Acan [113] combined handcrafted morphological features from the ECG signal with multi-stage features retrieved from different layers of a CNN in a late fusion strategy. Experiments showed that incorporating domain knowledge improves the classification results.

In [9], the authors explored ECG classification using GAF, RP, and MTF image transformations with two fusion strategies. The first, called multimodal image fusion, concatenates images from different transformations and classifies them through a single model, resembling early fusion. The second, multimodal feature fusion, extracts features from GAF, RP, and MTF images using separate AlexNet models. These features are combined using a gated fusion network and classified with a SVM. More recently, in [123], a novel DL model named Multimodal Multi-Instance Learning Neural Network (MAMIL) was introduced for ECG classification. MAMIL integrates original ECG signals with GAF images, applying multi-instance learning and an attention-based feature fusion mechanism. Although this approach improves classification, it requires substantial computational resources.

Regarding incorporating other modalities beyond ECG representations, Xiao et al. [348] proposed a fusion network that incorporates demographic features with the 12 leads of the ECG, although limited to age and gender. Other studies have also explored multimodal fusion in various contexts. Rastgoo et al. [261] developed a driver stress level classification system using ECG, vehicle, and contextual data. The system employed separate CNNs to extract features, which were then concatenated and classified using an LSTM, demonstrating superior performance to traditional feature-based models. Hammad et al. [122] proposed a biometric authentication system using ECG and fingerprint data. They compared sequential and parallel systems, with the parallel system combining CNN-based feature extraction and a Q-Gaussian SVM for classification. The multimodal system outperformed unimodal approaches, albeit with only a modest improvement of about 1%. Bhatti et al. [40], on the other hand, included Electrodermal Activity (EDA) signals that were combined with ECGs using cross-attention mechanisms for emotion recognition tasks. In the same task, Bota et al. [46] added respiration and Blood Volume Pulse (BVP) signal to the previous ECG and EDA. However, the authors did not use DL methods to

combine all sources. Features were extracted and combined in traditional ML models at feature and prediction levels.

5.4.3 Methodology

Building on the potential of multimodal AI techniques and recent advances in leveraging diverse interpretations of ECG signals for multi-label disease classification, we will explore the impact of combining different data representations. We utilize both raw ECG signals and image transformations, specifically GAF, RP, and MTF [9], to provide a richer knowledge extraction for the robust diagnosis of cardiovascular diseases. The PTB-XL dataset serves as the foundation for this investigation, allowing us to assess how these varied signal representations contribute to the task of multi-label classification.

To fully realize the potential of this multimodal approach, we apply different neural network architectures and fusion strategies outlined in the literature that integrate the strengths of both 1D signal data and 2D image representations [138]. Through this work, we seek to contribute new insights into how multimodal techniques can enhance the effectiveness of DL models for cardiovascular disease diagnosis.

5.4.3.1 Dataset

For this project, we utilized the PTB-XL open-source ECG dataset, published in 2020 through PhysioNet [331]. PTB-XL consists of 21,837 ECG records, each spanning 10 seconds and collected from 18,885 patients. This dataset includes 12-lead ECG recordings, annotated by up to two cardiologists, and is organized as a multi-label dataset. The diagnostic statements are aggregated into five primary categories, or super-classes, as detailed in Table 5.1. This table also highlights the number of records associated with each diagnosis, clearly illustrating the inherent imbalance present in the dataset.

Table 5.1: Description of records per label in the PTB-XL dataset, with the partition into the train, validation, and test sets, as suggested in [331], with the number of samples in each multi-label disease class.

Acronym	Super-class	No. Records			
		Train	Validation	Test	Total
AD	Normal ECG	7,254	916	913	9,083 (42.4%)
MI	Myocardial Infarction	4,389	544	553	5,486 (25.6%)
STTC	ST/T change	4,193	534	523	5,250 (24.5%)
CD	Conduction Disturbance	3,912	497	498	4,907 (22.9%)
HYP	Hypertrophy	2,121	271	263	2,655 (12.4%)
Total Samples		17,111	2,156	2,163	21,430

This work aims to classify each ECG record into one or more of the four main disease categories while treating the absence of any of these diseases as a Normal ECG. This

approach enables the models to specialize in recognizing each disease’s unique characteristics and patterns. A Normal ECG is implied when no positive classifications are made for any disease class. For a comprehensive evaluation, performance metrics for the Normal ECG/Absence of Disease (AD) class are also calculated, providing insights into how well the models can distinguish between diseased and normal conditions.

The dataset is balanced by sex (52% male and 48% female), with ages ranging from 0 to 95 years (median age of 62 and an interquartile range of 22 years). The data was collected at a 400 Hz sampling rate, followed by minor preprocessing steps such as spike removal and resampling. Although the PTB-XL dataset includes additional information, such as demographic details, diagnostic likelihoods, validation data, and signal metadata (e.g., signal quality), these supplementary features were not used in this study.

The PTB-XL dataset was developed to fill the gap in publicly available data for developing automatic ECG interpretation algorithms and has already been employed in several studies. As discussed in Section 5.4.2, DL models for ECG classification based on the 1D signals in this dataset have been proposed in various works [44, 296, 302, 329]. However, to our knowledge, approaches leveraging image-based and multimodal techniques had not been explored before using the PTB-XL dataset. This study aims to bridge that gap by investigating the potential of combining multiple data representations for more robust ECG classification.

5.4.3.2 Data Preprocessing

The PTB-XL dataset provides ECG data at two sampling frequencies: 100 Hz and 500 Hz, resampled from the original 400 Hz. Given that a sampling frequency of 90 Hz is recognized as the minimum needed to detect the distinct waveforms in the ECG signal (P, QRS, and T waves) [51], we opted to use the 100 Hz data, which balances performance and computational efficiency.

We applied a second-order Butterworth bandpass filter to preprocess the ECG signals. This filter had a high-pass cutoff frequency of 1 Hz to suppress baseline wander and a low-pass cutoff frequency of 45 Hz to eliminate high-frequency noise [185, 258, 295]. These settings were chosen to retain the essential frequency components of the ECG signal while filtering out irrelevant noise.

Given the computational costs of using all 12 ECG leads, we selected three leads: I, II, and V2. This choice was informed by the work of Zhu et al. [364], which demonstrated that these three leads contain most of the relevant information needed to reconstruct a 12-lead ECG.

For data partitioning, we adhered to the recommended training/validation/test split provided by Wagner et al. [331], utilizing the original data partitions. The authors suggested a 10-fold train-test split, achieved through stratified sampling while ensuring that all records from a given patient remained within the same fold to prevent data leakage. Notably, records in folds 9 and 10 had undergone at least one human evaluation, ensuring

higher label quality. Based on this recommendation, we used folds 1-8 as the training set, fold 9 as the validation set, and fold 10 as the test set.

Finally, we excluded any samples without ground truth labels, resulting in the final count available in Table 5.1.

5.4.3.3 1D Approaches

The input to the implemented 1D networks consisted of preprocessed ECG signals, represented as three sequences denoting the three used leads, each with a length of 1,000 samples (corresponding to 10 seconds of data recorded at 100 Hz). These sequences were used to train models with various architectures, including GRU and LSTM recurrent networks in both unidirectional and bidirectional configurations, i.e., the Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM). Additionally, we combined 1D CNN layers with RNN layers to leverage the strengths of both architectures. We also incorporated the self-attention mechanism detailed in Section 2.2.2 into the best-performing network to enhance performance. Figure 5.2 provides a general overview of the structure of these networks. The bidirectional models (BiGRU and BiLSTM) process the time series data in both forward and backward directions, capturing temporal dependencies from both perspectives.

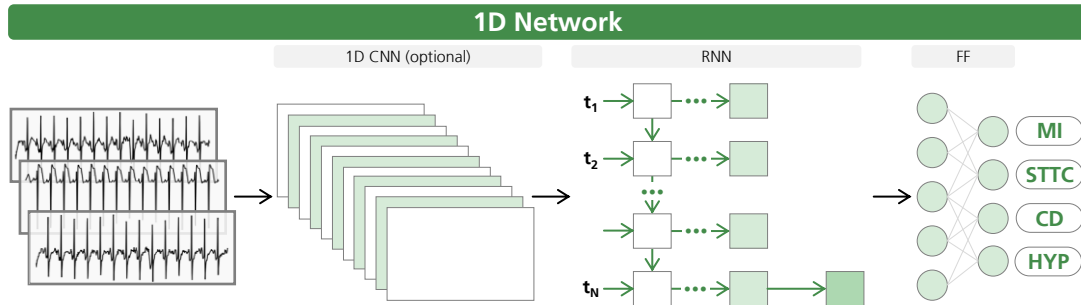


Figure 5.2: Overview of the proposed 1D network for ECG classification, based on an RNN and optionally a CNN.

5.4.3.4 2D Approaches

In image classification, CNNs such as AlexNet, VGGNet, and ResNet have demonstrated excellent performance across various applications, including ECG classification [55, 80]. We investigated these three architectures along with more modern approaches, such as MobileNetV2 [277] and AlexNet with a self-attention mechanism [326], to leverage recent advancements in DL.

For each ECG record, the three signals, each of size 1,000 samples, were transformed into images using three distinct methods: GAF, RP, and MTF [9]. These transformations resulted in images of size (3; 1,000; 1,000).

The conversion of a normalized ECG signal with n samples (s_k , for $k = 1, \dots, n$ and $0 \leq s_k \leq 1$) into the three image representations is carried out as follows:

- GAF: For each sample s_k , the angle $\beta_k = \arccos(s_k)$ is calculated. Each pixel (k, l) of the GAF image is then computed as $\cos(\beta_k + \beta_l)$.
- RP: Each pixel (k, l) of the RP image represents the Euclidean distance between two ECG samples (s_k and s_l).
- MTF: The signal is divided into B quantile bins, with each sample s_k assigned to the corresponding bin b_j ($j \in [1, B]$). A $B \times B$ matrix is built, where each entry w_{ij} represents the probability that a sample in quantile b_j is followed by a sample in quantile b_i . Each pixel (k, l) in the MTF image represents $w_{ij}|_{x_k \in q_i, x_l \in q_j}$. Following [9], B was set to 10.

By concatenating the images generated for each signal, we obtained an image of size (9; 1,000; 1,000) for each record. Due to memory constraints, these images were resized to (9; 256; 256) before being used to train the five CNN architectures. An overview of the proposed image-based classification framework is depicted in Figure 5.3.

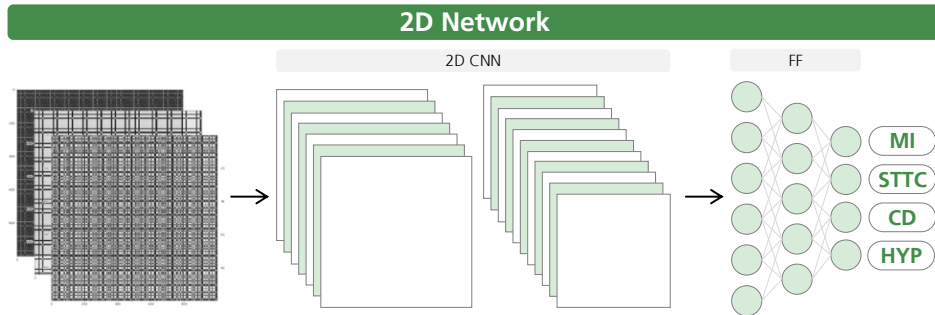


Figure 5.3: Overview of the proposed CNN-based framework for ECG classification.

5.4.3.5 Multimodal Fusion Approaches

In this work, we utilized multimodal fusion to combine structurally distinct representations of ECG data. The raw 1D ECG and its 2D representations are not common modality pairs, such as image and text, and neither are collected through distinct acquisition devices. However, they may provide networks with complementary information. By merging these representations, we aim to enhance the performance of classification models. Consequently, we explored the fusion of the best-performing unimodal networks. Three different fusion strategies were evaluated, each adopting distinct training approaches [138]. To further investigate the effectiveness of these strategies, we incorporated feature-level attention mechanisms across all fusion models, drawing on the attention frameworks proposed by Bahdanau et al. [20] and Vaswani et al. [326]. A self-attention layer was

integrated after the concatenation of unimodal features, enabling a comparison of the original fusion models with and without attention.

The first fusion method we tested was a late fusion approach, which combined the predictions from each network. As shown in Figure 5.4, this method concatenates the output predictions from the single-modality networks and passes them through a fully connected layer. The parameters of the individual networks were fixed, allowing the fully connected layer to transform the concatenated predictions into the final classification output.

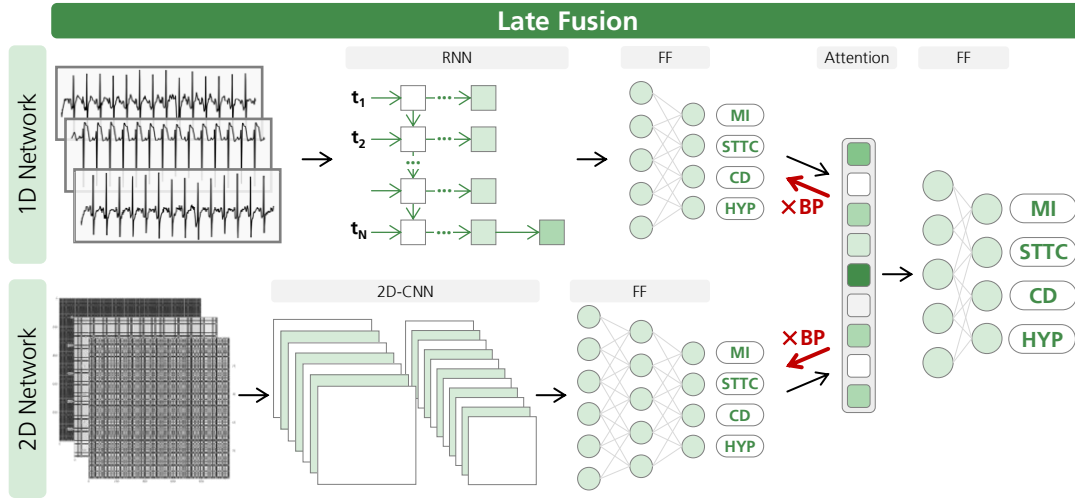


Figure 5.4: Overview of the proposed late fusion framework for ECG classification. BP represents the backpropagation to the unimodal networks, which does not occur in the late fusion strategy (red cross). Additionally, the effect of a self-attention layer after feature concatenation was tested against using solely a single fully-connected layer.

Next, we implemented an early fusion model that combined features extracted from the pre-trained single-modality networks, as illustrated in Figure 5.5. In this approach, activations from various layers of the unimodal networks were captured and fed into the early fusion network. While the activation from the last RNN block in the 1D branch was retrieved, activations from the third, fourth, and fifth convolutional blocks from the 2D branch were tested. To address the dimensionality mismatch between the 1D and 2D networks, we applied max pooling and a fully connected layer to the 2D network's output, reducing its feature size. The combined feature sets were then concatenated, followed by three fully connected layers to generate the final predictions. Each prior fully connected layer doubled the number of neurons.

Finally, we explored a joint fusion strategy to assess the potential benefits of fine-tuning the pre-trained parameters from the individual networks for capturing complementary information. As depicted in Figure 5.6, this approach involved retrieving activations from the last convolutional block or the first and second fully connected layers of the 2D network while always using the RNN block activations from the 1D network. Again, a fully connected layer was added after the 2D network to reduce the dimensionality and

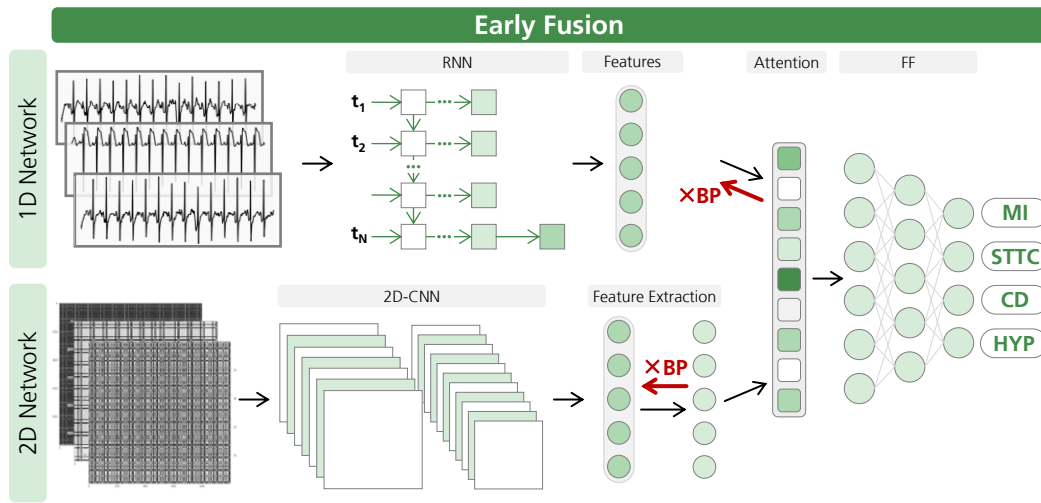


Figure 5.5: Overview of the proposed early fusion framework for ECG classification. BP represents the backpropagation to the unimodal networks blocked in early fusion (red cross). Additionally, the effect of a self-attention layer after feature concatenation was tested against using only the three fully connected layers scheme.

match the activation sizes from both unimodal networks. The fused features were then passed through three fully connected layers, including a double factor to each previous layer, which were trained from scratch for cardiovascular disease prediction. In addition, the performance of this joint fusion model was evaluated with and without the inclusion of a self-attention layer after the concatenation of the unimodal outputs.

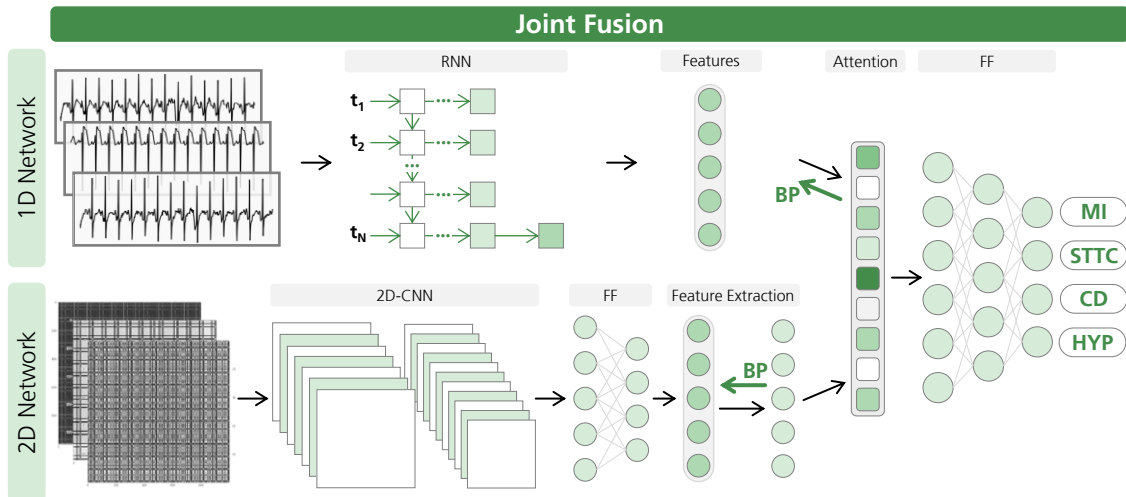


Figure 5.6: Overview of the proposed joint fusion framework for ECG classification. BP represents the backpropagation to the unimodal networks, which occurs in the joint fusion setting (green arrow). Additionally, the effect of a self-attention layer after feature concatenation was tested against using only the three fully connected layers scheme.

5.4.4 Experiments

To evaluate the performance of our proposed models across different modalities and fusion approaches, we conducted experiments using the PTB-XL dataset for multi-label classification of cardiovascular diseases described in Section 5.4.3.1. The experiments were designed to assess the effectiveness of 1D, 2D, and multimodal approaches in predicting the four primary diseases in the dataset: Myocardial Infarction (MI), ST/T change (STTC), Conduction Disturbance (CD), and Hypertrophy (HYP). Each of the four disease labels was considered a separate output node in the models, with the absence of any disease corresponding to the original NORM class. The models were optimized for multi-label classification, using a weighted binary cross-entropy loss to mitigate the impact of class imbalance:

$$\mathcal{L} = -\frac{1}{N} \sum_{o=1}^N \sum_{c=1}^M w_c [y_{o,c} \log(p_{o,c}) + (1 - y_{o,c}) \log(1 - p_{o,c})] \quad (5.1)$$

All models were trained to minimize the $\mathcal{L}$. $M = 4$ represents the number of diseases, y is the binary indicator of the ground truth class label for observation o , and p is the predicted probability of observation o belonging to class c , N is the total number of observations, and w_c is the class weight, calculated as inversely proportional to the number of samples of that class in the training set.

To evaluate the robustness of the trained models, we computed several performance metrics, including sensitivity, specificity, precision, accuracy, and G-Mean. G-Mean was selected as the primary metric for model selection, given its balanced consideration of sensitivity and specificity:

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (5.2)$$

The decision thresholds for each disease label were optimized on the validation set to maximize the G-Mean. Models were evaluated on the test set based on the parameters saved at the epoch with the lowest validation loss.

Before training the model, a hyperparameter tuning phase was implemented for each strategy. The values tested are detailed in Table B.1. Afterward, both the 1D and multimodal fusion models were trained for a maximum of 200 epochs, while 2D models were capped at 100 epochs. An early stopping mechanism prevented models from overfitting with patience of 10 epochs.

5.4.5 Results

The implemented approaches for 1D and 2D ECG classification were trained with the described hyperparameters to select the best-performing strategies for each modality. These were then used as the 1D and 2D branches to optimize the multimodal networks, thus providing a fair way to compare results. The optimal hyperparameters for each trained network are detailed in Table B.2.

5.4.5.1 1D Approaches

For each of the six 1D architectures described in Section 5.4.3.3, hyperparameters were optimized based on validation loss, with the best configurations selected for testing on the held-out test set. The performance metrics for these models are summarized in Table 5.2, and results for each metric are available in Table B.3.

Table 5.2: 1D ECG classification results obtained with RNN and CNN-RNN models, in percentage (%). The best results are highlighted in bold.

Model	Parameters	Sensitivity	Specificity	G-Mean
GRU	249,732	79.67	81.04	80.35
GRU Attention	249,861	76.55	80.15	78.33
LSTM	794,628	75.42	80.09	77.72
BiGRU	399,620	78.95	81.07	80.00
BiLSTM	532,484	78.18	80.87	79.51
CNN-GRU	219,536	72.33	78.35	75.28
CNN-LSTM	81,296	74.04	77.06	75.53

Among the models, the GRU architecture achieved the highest sensitivity, with a 79.67%, while the BiGRU model attained the best specificity, at 81.07%. When considering the G-Mean of sensitivity and specificity, an important measure for balancing the detection of both positive and negative cases, the GRU model emerged as the top performer, with a G-Mean of 80.35%.

Interestingly, adding 1D convolutional layers before the recurrent layers did not lead to improvements in model performance, suggesting that the combination of CNN and RNN architectures did not significantly benefit this specific task. Additionally, GRU-based models consistently outperformed their LSTM counterparts in terms of G-Mean while also being more parameter-efficient. Introducing bidirectionality did not enhance the performance of the GRU model, although it did slightly improve results for the LSTM architecture. Furthermore, including a self-attention mechanism in the GRU model did not yield any notable gains in performance. This indicates that the attention mechanism did not add significant value in this context, potentially related to the temporal continuity between values, which attention may fail to capture or overfitting to the train set.

5.4.5.2 2D Approaches

Regarding the 2D convolutional models described in Section 5.4.3.4, Table 5.3 presents the disease-aggregated results, while B.4 presents the performance metrics per disease.

Based on the G-Mean, the best overall performance was achieved by the AlexNet with Attention (AlexNetAtt) model, with a G-Mean of 75.95%. Additionally, VGG16 recorded the highest sensitivity at 74.62%, while MobileNetV2 and AlexNetAtt shared the highest specificity at 81.98%. Given the importance of balancing sensitivity and specificity in

5.4. IMPROVING CARDIOVASCULAR DISEASE DIAGNOSES USING ELECTROCARDIOGRAMS

Table 5.3: 2D ECG classification results obtained with CNN models, in percentage (%). The best results are highlighted in bold.

Model	Parameters	Sensitivity	Specificity	G-Mean
AlexNet	46,334,468	73.85	76.38	75.11
ResNet	1,494,356	67.31	76.48	71.75
VGGNet	285,347,124	74.62	75.77	75.19
MobileNetV2	2,230,724	65.05	81.98	73.03
AlexNetAtt	46,335,880	70.36	81.98	75.95

clinical contexts, the AlexNetAtt model stands out as the most effective choice for ECG classification using these image-based transformations. Notably, AlexNetAtt also offers an advantage in terms of model complexity, requiring fewer parameters than the VGG16 architecture.

These findings demonstrate that the ECG-to-image transformations proposed by Ahmad et al. [9], originally applied to single-heartbeat ECG signals, can be successfully extended to sequences of multiple heartbeats. The resulting images capture both the temporal characteristics and relative amplitude information of the ECG signal’s various waveforms, allowing CNNs to effectively extract features for accurate ECG classification.

5.4.5.3 Multimodal Fusion Approaches

Table 5.4 displays the test set results for each fusion strategy, both with and without a self-attention layer. In addition, Table B.5 presents extended results per disease. The best-performing unimodal networks, the GRU-based model for the 1D ECG signal and the AlexNetAtt model for the 2D ECG representations, were combined to form these multimodal fusion networks. For each strategy, the model with the lowest validation loss was selected for final evaluation.

Table 5.4: Multimodal fusion ECG classification results, in percentage (%). The parameters in parentheses represent the total number of parameters in the model, while the main value refers to the trainable parameters. The best results are highlighted in bold.

Model	Parameters	Sensitivity	Specificity	G-Mean
Late Fusion	1,668 (46,587,278)	79.82	79.02	79.42
Late Fusion Attention	1,884 (46,587,494)	81.53	67.80	74.35
Early Fusion	623,620 (1,057,032)	77.93	79.68	78.80
Early Fusion Attention	592,516 (878,216)	76.87	75.54	76.20
Joint Fusion	4,137,480	73.67	79.55	76.56
Joint Fusion Attention	40,712,200	74.62	79.79	77.16

Among the fusion strategies, late fusion without the self-attention layer achieved the highest overall performance, with a G-Mean of 79.42%. However, this model did not

yield the best individual metrics: the highest sensitivity (81.53%) was recorded by the late fusion model with self-attention, and the highest specificity (79.79%) was obtained with the joint fusion model that also incorporated attention.

Interestingly, these results challenge expectations regarding model complexity. Despite the increased parameters and access to latent space information from earlier network layers, neither the early fusion nor the joint fusion approaches outperformed the simpler late fusion model. This suggests that combining output probabilities from single-modality networks proved more effective than fusing intermediate activations.

The impact of the self-attention layer was also limited. While it improved performance in the joint fusion strategy, this approach significantly increased the model's complexity, resulting in a tenfold increase in the number of parameters compared to other models. Adding self-attention in the other fusion strategies led to decreased performance, indicating that it may not always contribute positively to multimodal learning in this context.

Furthermore, the results indicate that fusion models did not significantly enhance performance over the unimodal networks. Contrary to expectations, these findings suggest that multimodal DL may offer limited advantages for ECG classification tasks, raising questions about the effectiveness of combining different ECG representations in this domain.

5.4.5.4 Comparison

The multi-label classification of cardiovascular diseases was approached through three distinct strategies: using the original ECG signal, leveraging signal transformations into GAF, RP, and MTF images, and combining both in multimodal fusion strategies. These different data schemes were employed to assess the relevance of image transformations and multimodal fusion in enhancing cardiovascular disease prediction. Additionally, we explored various DL techniques, incorporating attention-based strategies to identify the most effective approaches.

Table 5.5 summarizes the key performance metrics for the best-performing networks in each data scheme: the GRU model in the 1D time series sub-net, the AlexNetAtt model in the 2D image sub-net, and the late fusion strategy among the multimodal fusion networks.

5.4.6 Discussion

This work aimed to evaluate the potential of combining different ECG representations to improve the robustness of cardiovascular diseases classification in an opportunity to go deeper into leveraging multimodal AI for better clinical DSSs. We compared multiple DL approaches exploring both 1D and 2D representations from ECG signals, initially within independent models and later combining them in multimodal fusion networks. A key objective was to determine whether merging different representations of the same signal could improve classification performance since previous studies suggested that distinct

5.4. IMPROVING CARDIOVASCULAR DISEASE DIAGNOSES USING ELECTROCARDIOGRAMS

Table 5.5: Final ECG classification results for best-performing models, in percentage (%). The parameters in parentheses represent the total number of parameters in the model, while the main value refers to the trainable parameters. The best results are highlighted in bold.

	MI	STTC	CD	HYP	AD	Total
<i>1D Approach - GRU</i>						
Sensitivity	84.81	83.37	78.71	72.24	77.11	79.67
Specificity	78.82	83.96	86.31	71.74	87.20	81.04
G-Mean	81.76	83.66	82.42	71.99	82.00	80.35
Precision	57.90	62.37	63.23	26.13	81.48	58.90
Accuracy	80.35	83.82	84.56	71.80	82.94	80.69
<i>2D Approach - AlexNetAtt</i>						
Sensitivity	60.94	74.76	59.84	58.56	82.69	70.36
Specificity	77.45	83.66	88.11	82.68	76.40	81.98
G-Mean	68.70	79.08	72.61	69.58	79.48	75.95
Precision	48.14	59.33	60.10	31.88	71.90	57.11
Accuracy	73.23	81.51	81.60	79.75	79.06	79.03
<i>Multimodal Approach - Late Fusion</i>						
Sensitivity	75.59	88.91	83.73	77.19	76.34	79.82
Specificity	86.27	77.56	83.18	64.11	88.72	79.02
G-Mean	80.75	83.04	82.96	70.34	82.30	79.42
Precision	65.41	55.82	59.54	22.94	83.17	56.47
Accuracy	83.54	80.31	83.08	65.70	83.50	79.22

signal representations capture different aspects of the data, potentially leading to better overall performance.

From the results in Table 5.5, it is clear that the 1D approach consistently outperforms both the 2D and multimodal fusion strategies regarding overall performance, particularly in the G-Mean. A plausible explanation for this result lies in the 1D models' ability to directly process raw ECG signals, effectively capturing intricate temporal patterns that may be lost during image-based or combined data representations. The inherent complexity of integrating modalities in multimodal approaches could have contributed to their relatively lower performance. Challenges such as the dominance of one modality, increased model complexity leading to overfitting, or inefficiencies in how different modalities complement each other might have played a role in this outcome. This fact shows that the technical challenges of multimodal AI discussed in Section 5.2 are still open for further research.

Moreover, the simplicity of the 1D models provides a practical advantage. These models do not require an additional signal-to-image transformation step and have fewer parameters (Table 5.2) compared to their 2D counterparts (Table 5.3). While these findings may initially cast doubt on the utility of multimodal DL in ECG classification, it is important to consider that multimodal fusion still showed its strengths, particularly in boosting sensitivity, which is crucial in clinical settings. Detecting diseases accurately is of utmost importance, even if it comes at the cost of increased false positives. In clinical

practice, missing a diagnosis can have far more severe consequences than flagging a false positive. Nevertheless, while this work focused on a hypothetical scenario using a public dataset, a real-world definition of costs for determining the performance priorities should be considered in case of delivering such a DSS to clinical practice.

Despite the potential for combining complementary information from different modalities, the multimodal fusion strategies did not lead to a significant improvement. The late fusion approach without a self-attention layer yielded the highest G-Mean (79.42%) but did not consistently surpass the best 1D model results. The complexity introduced by multimodal fusion, including the need to optimally balance information from different modalities, may have hindered its potential benefits. The fusion of output probabilities in late fusion outperformed other methods, possibly due to the simplicity of combining already optimized unimodal predictions. Additionally, including a self-attention layer did not lead to substantial improvements, except in the joint fusion approach, which marginally improved results but at a significant computational cost. This suggests that while attention mechanisms have proven beneficial in other tasks, their impact on multimodal ECG classification may be limited. Solutions for these limitations could include exploring weight-based modality combinations, either defined in shallow terms or dynamically adjusted per sample, or end-to-end training of the multimodal networks, especially the less computationally intensive late fusion strategy.

The multimodal approach showed particular strengths when examining individual disease performance, especially in detecting STTC with high sensitivity (88.91%). However, no single strategy consistently outperformed the others across all diseases. For example, the 1D models performed best in detecting MI with a sensitivity of 84.81%. The multimodal approach excelled in sensitivity but struggled with maintaining high specificity, while the 2D models lagged in both metrics. Across all strategies, precision remained relatively low, highlighting the challenge of balancing false positives and false negatives. This is a critical consideration in clinical contexts, where high precision is desirable to avoid unnecessary treatments, but not at the expense of sensitivity.

5.4.6.1 Literature Comparison

Previous studies on the PTB-XL dataset achieved average accuracy and sensitivity ranging from 66.2% to 93.3% [296, 297, 329, 351]. Our best-performing 1D and multimodal approaches improved over some of these results, achieving average accuracy and sensitivity of 80.69%/79.67% and 79.22%/79.82%, respectively. While Virgeniya et al. [329] reported higher sensitivities and specificities, their results were significantly enhanced by synthetic data generation, which was not used in our study. Additionally, our models demonstrated strong performance without manual feature extraction, making them more generalizable.

Yang et al. [351] achieved slightly higher sensitivities but at a higher computational cost. Our models, particularly the GRU network, were more efficient, with fewer parameters, highlighting a balance between performance and resource use, crucial for practical

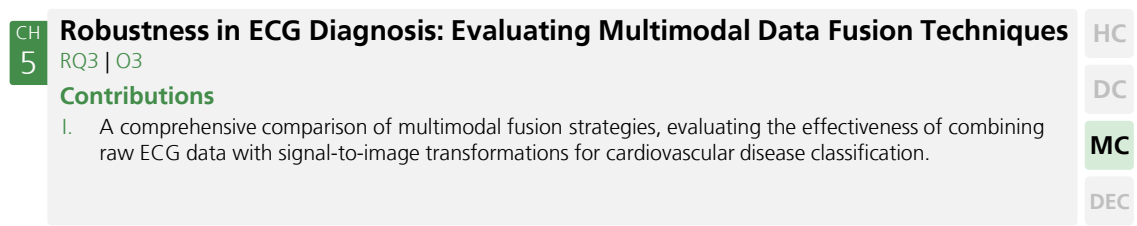


Figure 5.7: Overview of Chapter 5 contributions, aligned with the respective approaches of the holistic framework proposed in Chapter 3.

deployment in clinical settings.

Our study shows that while multimodal DL may offer specific advantages in certain contexts, particularly in improving sensitivity, the simpler 1D models remain highly effective and reliable for ECG classification tasks. Future work could explore additional strategies to improve the balance between sensitivity and specificity, particularly synthetic data augmentation.

5.5 Final Remarks

In our pursuit of building more robust models for real-world clinical systems, we began by focusing on cardiothoracic surgery in Chapter 4. There, we explored the integration of domain-specific information to enhance risk prediction models and investigated the potential of correlation-based missing data imputation techniques. These efforts highlighted the critical role of expert-driven approaches in improving model performance, but they also revealed limitations. Our work was restricted to structured clinical variables and feature extraction processes, and we found that human-centered strategies, while valuable from design to application, often suffer from limited generalization.

Building on these findings, we explored a different use case: the multi-label classification of cardiovascular diseases from ECGs. The goal was to investigate whether incorporating multiple representations of ECG data in a multimodal approach would improve model robustness. This allowed us to study the impact of multimodal AI in clinical DSSs.

In this chapter, we explored the concept of multimodal AI, defining a modality and examining the technical challenges addressed in the literature. We focused specifically on fusion techniques through different strategies for merging raw ECG data with signal-to-image transformations that have proven useful in previous research [9]. After optimizing the unimodal models, we trained and tested various multimodal fusion strategies. The summary of contributions from Chapter 5 is shown in Figure 5.7.

Our findings indicated that contrary to our expectations, combining both modalities did not significantly enhance overall classification performance for cardiovascular diseases. However, the multimodal approach did lead to improved sensitivity in detecting certain conditions, which is crucial in clinical contexts where avoiding false negatives is a priority.

This suggests that while multimodal AI techniques offer some advantages, particularly in enhancing sensitivity, they may not always result in significant overall performance gains. We also recognized that multi-label models bring computational efficiencies, but individual disease models or ensemble strategies could potentially increase robustness in real-world DSSs. Leveraging uncertainty measures, as proposed by Barandas et al. [24], could further improve decision-making by determining which models to apply for each disease or even for individual predictions.

Our exploration of multimodal AI techniques through different representations of ECG data highlights the potential for increasing model robustness. However, actual performance gains may come from incorporating more traditional multimodal data sources, such as demographics or clinical history. Much of the existing research in multimodal AI focuses on combining images and text, often outside the medical domain (e.g., CLIP [255], DALL-E [259]), though some approaches are specifically designed for medical data, such as LLaVA-Med [183]. These prompt-based strategies are designed for specific tasks like image segmentation, which differ technically from the more comprehensive and dynamic demands of traditional DSSs in healthcare. In addition, the challenges of handling missing modalities, as discussed by Ma et al. [204], also remain an open area of research with considerable potential for impact.

Recognizing the limitations of traditional clinical DSSs, which often struggle with incomplete data across multiple modalities and patient histories, we shifted our focus to the recent breakthroughs in large, general models that have significantly advanced AI in the past few years. While these models, particularly those developed in the NLP domain, have demonstrated impressive capabilities across various tasks, their application to clinical data remains largely unrealized. In the next chapter, we will explore how these large models can be adapted to address the challenges of clinical DSSs, aiming to build a system that fully leverages the complexity of available patient data while generalizing across diverse healthcare scenarios. Our goal is to develop a robust framework that can handle sparse and longitudinal data without depending on traditional preprocessing techniques like imputation or resampling, marking a significant step toward more adaptable and resilient clinical DSSs for the future.

ROBUSTNESS IN CONTINUOUS PATIENT MONITORING: LEVERAGING ASYNCHRONOUS INPUTS AND TRANSFORMERS

6.1 Introduction

The previous chapters have highlighted several challenges inherent to using AI-based DSSs in healthcare. In Chapter 4, dealing with real-world cardiothoracic surgery data revealed issues of data heterogeneity and inconsistency, while our exploration of ECG multi-label classification in Chapter 5 studied the potential for multimodal DSSs by combining different interpretations of data.

However, beyond heterogeneity, inconsistency, and multimodality, the critical challenge of dealing with the longitudinal nature of medical data remains. Medical information is collected throughout every patient's life, forming a comprehensive but underexplored "biological biography." This longitudinal data presents immense potential but also introduces limitations that current methods struggle to address [59].

The complexity of medical data collection is compounded by the fact that it is often asynchronous [318]. Each patient interaction with the healthcare system is driven by different reasons, such as routine checkups, emergent symptoms, or ongoing treatments, leading to the collection of various data modalities at irregular intervals. This results in a sparse, patient-specific data matrix that is challenging to process. While we have previously studied the impact of missing data and developed imputation techniques, such methods fail to address the inherent inconsistencies and asynchronous nature of longitudinal medical data.

ICU environments present these challenges and underscore the need for robust DSSs in critical care settings. Patients in ICUs are often in life-threatening conditions, either due to severe illness or high-risk surgical procedures. In these situations, effective risk prediction or forecasting of critical events could enable clinicians to act preemptively,

potentially preventing adverse outcomes and saving lives. Yet, despite the intensive and continuous monitoring in ICUs, the resulting data is highly granular but also sparse and irregular. While some data points, such as laboratory results or intake records, may follow a predictable pattern, others are sporadic.

Applying traditional missing data imputation methods to such a data structure could create a "clean" data matrix that fails to reflect reality accurately and adds unnecessary complexity to the system. The irregular nature of ICU data requires new approaches to handle these unique challenges without relying on conventional preprocessing.

Recent advancements in NLP have introduced powerful attention-based models demonstrating remarkable performance and generalization capabilities [235]. However, despite their promise, these breakthroughs in text processing have not yet been fully translated into the complexities of medical data. Some studies have attempted to transform medical data into text [139], while others have focused on a limited set of variables [182]. Still, none have yet managed to harness the full scope of available clinical data. Nevertheless, attention-based techniques offer significant potential for handling sequences of clinical data points, potentially bypassing the need for traditional heavy preprocessing. By processing clinical data in its entirety and incorporating the full complexity of each measurement, these techniques can potentially revolutionize healthcare through more robust DSSs.

This chapter will explore these recent methodologies and their applicability to medical data, particularly in ICU settings. We will assess the limitations of existing approaches and employ data-centric and especially model-centric approaches to propose innovative strategies to address these gaps. The ICU use case will serve as the foundation for this investigation, with the ultimate goal of developing adaptable and effective DSSs that can be generalized from a large ICU model to other clinical scenarios.

6.2 Fundamentals of Longitudinal Data

Longitudinal data consists of the collection of repeated observations from the same individual over time [102]. This data characteristic spans several domains, but it is essential in medical settings for tracking patients' health progression, especially in environments like the ICU or chronic disease monitoring. Longitudinal data is valuable because it captures temporal dynamics, allowing clinicians to observe how variables change over time in response to treatments or disease progression [59].

Modeling longitudinal medical data presents great potential for useful DSSs that assist clinicians in processing historical patient data. Diseases evolve and originate from pathophysiological mechanisms over time. While models that rely on data from a single moment can be helpful in tasks such as detecting a mass in an imaging exam or diagnosing a cardiovascular disease from an ECG, some tasks benefit from leveraging a more extended patient history, such as predicting the evolution of neurodegenerative diseases.

6.2.1 Challenges

Longitudinal data is essential for understanding evolving patterns, predicting future outcomes, and planning effective interventions based on the patient's clinical history. However, it presents several inherent challenges and complexities.

One of the defining characteristics of longitudinal data is temporal dependency. Observations collected at different time points are not independent since the values recorded earlier can affect or be predictive of those recorded later [59]. This dependency breaks the "independent and identically distributed" assumption between data points, requiring models that appropriately account for the correlation across different time points while handling data sequences of varying lengths [59].

Another challenge is missing or asynchronous data points, which occur when some observations are skipped or collected at irregular intervals. In Chapter 4, we discussed the sources and implications of missing data, particularly in healthcare. Unrecorded clinical interactions, distinct intervention strategies, or problems in data registration exacerbate when dealing with longitudinal medical data [59]. While missing data can generally be considered an occasional event, asynchronicity is a characteristic of medical data standards and clinical practice. For example, vital signs may be measured continuously, while lab tests are only conducted periodically. Leveraging traditional missing imputation strategies may introduce bias in data and affect AI-based DSSs robustness [304].

Additionally, longitudinal data is often heterogeneous and complex, combining multiple data modalities, such as multivariate physiological signals, structured lab results, and unstructured clinical notes [59]. High dimensionality is a related characteristic, where many variables, such as symptoms, treatments, and test results, are monitored concurrently, leading to challenges in feature selection and computational complexity.

This heterogeneity makes developing a unified approach to handling diverse data types more difficult, especially when data is collected asynchronously. Combining all factors creates highly complex data vectors where each patient's medical history is unique.

6.2.2 Longitudinal Modeling

In healthcare, longitudinal models aim to track a patient's health trajectory over time by learning the temporal dynamics of medical data. Longitudinal modeling has been an intense field of study, with different strategies proposed and distinct technical approaches.

In clinical contexts, survival models are commonly used to predict time-to-event outcomes, e.g., time until ICU mortality or time until the next complication. These models estimate the probability of an event occurring at a specific time and account for censored data, where longitudinal data is available for some patients who have not yet registered the particular event. The Kaplan-Meier estimator and Cox proportional hazards model are classic survival models used for longitudinal data [19, 82].

Following improvements in the available computational infrastructure to support the usage of DL, sophisticated RNNs such as LSTMs [133] and GRUs [67] have been

increasingly applied to longitudinal data. These models are designed to capture temporal dependencies and have demonstrated strong performance in handling time-series data. For instance, in critical care settings, RNNs can monitor vital signs and laboratory results in real-time, predicting patient deterioration or alerting clinicians to early signs of sepsis. However, a limiting factor of such strategies is the common incapacity to deal with asynchronous variables, often requiring missing imputation [224].

Attention-based models, particularly transformer architectures, initially introduced by Vaswani et al. [326], have also emerged as powerful tools for longitudinal data analysis. These models leverage self-attention mechanisms to focus on essential elements within the data, allowing them to handle long-range dependencies. Self-attention and transformers have proven effective when it is necessary to weigh time points differently depending on their relevance to the current prediction [156].

The effective use of longitudinal modeling techniques in healthcare has the potential to significantly improve patient outcomes in diverse DSSs. However, these models must be carefully designed to address the inherent challenges of longitudinal data, including missingness, asynchronicity, and heterogeneity, while being scalable to accommodate large and diverse patient populations.

6.3 Transformer Models

The transformer architecture, introduced by Vaswani et al. in the paper "Attention is All You Need" [326], revolutionized the field of ML by replacing traditional recurrent and convolutional networks with a novel mechanism based on self-attention. Unlike previous models that processed data sequences serially, such as word by word in a sentence or point by point in a signal, transformers allow parallelization. This architecture explores multiple combinations of pairs of words and data points at the same time, making them highly efficient for processing large datasets.

At the core of the transformer model is the self-attention mechanism, detailed in Section 2.2.2, which allows the model to weigh the relevance of different words or data points relative to each other, regardless of their position in the sequence. This mechanism enables transformers to capture long-range dependencies, crucial in tasks like language modeling, machine translation, and time-series analysis [326].

Transformer models are currently the state-of-the-art technology to develop general models with impressive performance in multiple tasks. Large Language Models (LLMs) focus on processing text data and present three key features: large model size, scaling to billions or trillions on parameters; large generalization, given the straightforward adaptation to multiple tasks; and large-scale pre-training, using immense non-labeled data in self-supervision schemes [250]. Two of the most prominent transformer-based models are the Bidirectional Encoder Representations from Transformers (BERT) and the GPT, each designed with distinct architectures and purposes.

BERT, introduced by Devlin et al. from Google in 2019 [76], is designed as a bi-directional model. It processes text by looking at both the left and right context simultaneously, allowing it to understand a word in the full context of a sentence. BERT has demonstrated effectiveness in tasks that require a deep understanding of text, such as question answering, sentiment analysis, and named entity recognition. BERT's training process includes two primary tasks: Masked Language Modeling (MLM), where words in a sentence are randomly masked, and the model is tasked with predicting them, and Next Sequence Prediction (NSP), where it predicts whether two sentences appear sequentially in the text.

GPT, introduced by OpenAI in 2018 [254], is an auto-regressive model, i.e., processes text from left to right and predicts the next word in a sequence based on the previous words. Unlike BERT, GPT excels in text generation tasks and is used primarily for text completion, translation, summarization, and other tasks where generating coherent, contextually appropriate text is the goal. GPT's pre-training focused instead on predicting the next token in a sequence, which makes it more suited for generation tasks. Since the first GPT, OpenAI has been introducing new versions of the model, exploring multimodal features with larger parameter sets and combinations of smaller models in a mixture of experts setting, although with limited disclosure of actual models characteristics [235].

Transformer models like BERT and GPT are often referred to as foundational models because they are pre-trained on massive amounts of data and can be fine-tuned for a variety of downstream tasks with smaller annotated datasets [349]. This approach is rooted in transfer learning, where a model trained on one task is adapted to perform well on another task without requiring extensive training from scratch [256].

Representation learning is an important characteristic for this purpose [32]. During pre-training, transformers learn to generate dense, contextualized embeddings, i.e., representations, of the input data. These embeddings capture both semantic and syntactic information, allowing models to generalize across tasks.

Fine-tuning pre-trained models involves adjusting these representations to optimize performance on a specific task, such as predicting a particular outcome or forecasting future values. Through transfer learning from transformer models, clinical DSSs may achieve higher performance with less labeled data, addressing challenges related to data scarcity [250].

6.4 Developing Transformer Models for Critical Care

6.4.1 Objectives and Significance

Analyzing complex, multivariate, and asynchronous data presents significant challenges across many fields beyond healthcare, such as finance or retail. In healthcare, real-world data often contain missing values, inconsistent measurement frequencies, and non-linear relationships influenced by hidden factors. Traditional models struggle to

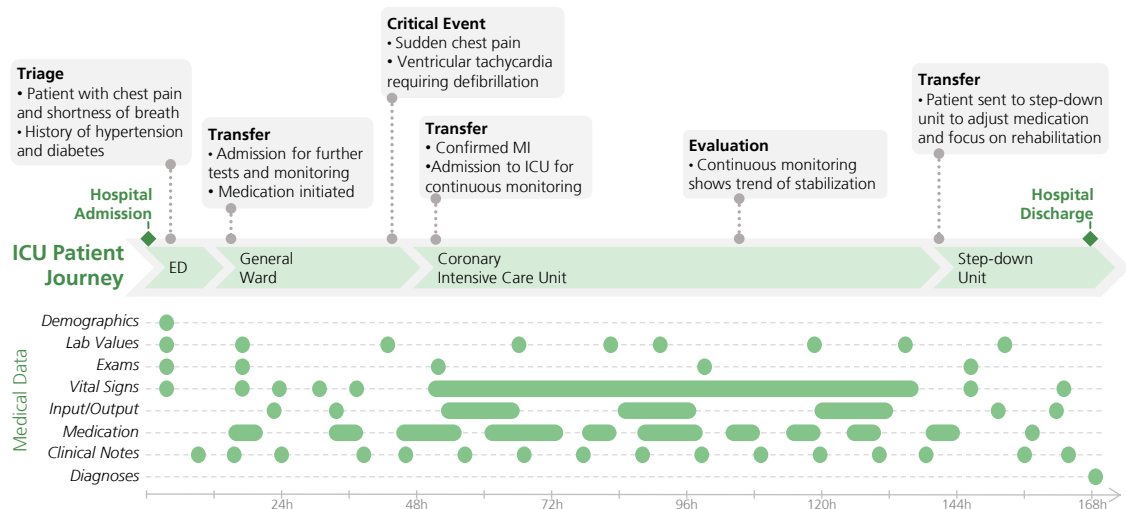


Figure 6.1: Example of an ICU patient journey and generated data, which is richer and more intensive in critical care. ED refers to the Emergency Department.

handle these complexities, particularly regarding irregular timing between observations and the interdependencies of multiple variables. These challenges become especially acute in ICUs, where patients are in critical condition, and robust DSSs are essential for rapid and informed decision-making.

ICUs continuously generate high-fidelity data from monitoring equipment and patient management systems [151], as depicted in Figure 6.1. This data is not only vast but deeply interconnected. Changes in one measurement, such as blood pressure, often signal changes in others, like kidney function. ICU data also vary in temporal resolution. While heart rate may be recorded every few minutes, lab tests might only be conducted daily [149]. Moreover, the data is inherently multimodal, combining structured inputs from EHRs, unstructured notes, and semi-structured machine outputs [216, 257]. This heterogeneity presents significant challenges for predictive models, particularly in maintaining integrity in the face of imbalanced target distributions, where normal conditions vastly outnumber critical events [263].

Large ICU databases like the MIMIC-IV [112, 151] have spurred research into applying ML to critical care, but existing limitations hinder real-world deployment of effective DSSs [84]. Traditional models are often hampered by the need to preprocess data into structured formats [220]. Techniques such as interpolation, imputation, and resampling risk distort the true nature of the data. More sophisticated models, including BEHRT [186], MedBERT [260], and DuETT [174], have improved data representation by producing richer embeddings from sparse longitudinal data. However, these models tend to focus on narrow data sources like ICD codes [132] or limited clinical variables, which constrains their ability to fully utilize the comprehensive data streams available in ICU settings. ICD codes, for instance, are primarily used for billing purposes and are often recorded after discharge, failing to provide timely medical insights.

In response to these challenges, in this section, **we introduce ICU-BERT, a transformer-based architecture designed to handle the unique complexities of ICU data.** ICU-BERT leverages the foundational BERT architecture [76] to process medical data sequentially, with each registry in a patient’s medical record representing a data point, i.e., a token in the input sequence. This model accommodates the sparse, multivariate, and irregular nature of ICU data, improving data representation and predictive performance across varied clinical settings. Unlike previous models, ICU-BERT introduces a multi-token input strategy that more effectively captures intricate medical data details, using a quadruplet representation of medical records to interpret clinical context better.

By incorporating pre-trained textual embeddings, ICU-BERT enhances its ability to generalize across different clinical settings. These embeddings provide a sophisticated initial representation of medical concepts, allowing the model to adapt to various ICU data streams. To optimize the learning of complex relationships, ICU-BERT employs a novel pre-training masking task coupled with a multi-task learning loss, improving the model’s robustness in real-world tasks. Pre-trained on the extensive MIMIC-IV database, ICU-BERT is well-equipped to handle the diverse and intricate nature of ICU data, offering a flexible and highly adaptable approach to clinical decision support.

We validated ICU-BERT across multiple real-world tasks, demonstrating its ability to generalize effectively to additional ICU datasets. Fine-tuning was performed on five diverse tasks from DuETT and the Yet Another ICU Benchmark (YAIB) framework [335], spanning classification, regression, and continuous monitoring tasks. Furthermore, ICU-BERT was evaluated on MIMIC-IV as well as other datasets like eICU [243], HiRID [95], and MIMIC-III [150]. In several cases, ICU-BERT not only matched but also surpassed existing models, underscoring its potential to enhance decision support systems in critical care significantly.

6.4.2 Literature Review

Using ML to process extensive longitudinal health data has been a long-standing effort beyond the context of critical care [340]. Medical records are inherently longitudinal, continuously accumulated throughout an individual’s life. Each visit to a health center generates new data points, including clinical notes, ICD codes, laboratory tests, exams, and prescribed medications. In ICUs, data collection is even more granular, with measurements taken by various equipment at minute or second intervals. Multiple predictive tasks have been studied in the literature aiming to optimize ICU resources. Predicting hospital, ICU and post-discharge mortality are often targeted, similarly to the cardiothoracic surgery use case of Chapter 4 [184, 236]. The forecasting of sepsis, Acute Kidney Injury (AKI), readmissions, and more recently, COVID-19 outcomes are also often found in the literature [220, 308].

Traditionally, ML-based DSSs require intensive data processing and feature extraction [171]. With the advancements in DL, recurrent models were employed to capture temporal

relationships from longitudinal data [107]. Scherpf et al. [283] proposed an RNN for predicting sepsis with nine parameters, and Kessler et al. [162] used 14 clinical features in an LSTM for predicting readmission to the ICU. Despite the potential of RNNs to handle temporal data, they lack effectiveness for sparse and asynchronous ICU data due to their need for fixed temporal resolution, inability to handle long data series, and reliance on a fixed variable matrix.

Later, HE-LSTM [194] introduced an event gate to model irregular intervals, and T-LSTM [224] adjusted for elapsed time, enhancing the modeling of sparse temporal data. Che et al. [62] proposed GRU-D, an RNN that incorporated a missing data matrix in addition to the time interval information. SAnD [301] was the first model based on self-attention, which employs a convolutional embedding layer to manage the input vector size. Also employing attention and a missing value mask, Yu et al. [354] proposed a GRU network to predict hospital mortality, assisted by a value reconstruction decoder to help regularization.

Recent advances in transformers have led to models such as BEHRT [186], which treats diagnosis codes as tokens within a limited vocabulary of 301 codes, modeling hospital visits as sentences and integrating age with position and segment embeddings. This approach utilized a MLM task for pre-training, enabling predictions of future disease codes. Despite its restricted scope, BEHRT was the first transformer for medical data to gain traction in the literature.

Models like Med-BERT [260] and ExMed-BERT [182] have then focused on different applications and incorporating more information, with the latter applying late fusion to include continuous data. Other models like Rare-BERT [246] and life2vec [282] expanded the vocabulary to cover a broader array of data types. ExBEHRT [272] and BRLTM [215] added demographic and lab data information through separate embedding layers. Hierarchical schemes such as Hi-BEHRT [187] and graph-based approaches such as GT-BEHRT [244] have been proposed to manage large input sequences by processing sets of visits before integrating them into a complete patient profile.

Using input vocabularies often limits data richness, as continuous values must be categorized. STraTS [318] addresses this with a triplet embedding scheme for features, continuous values, and time. In contrast, DuETT [174] uses a dual attention strategy to capture time- and event-based representations.

Another limitation is the restrictive input window size of architectures based on BERT. Fallahpour et al. [94] recently proposed EHRMamba, a model developed using the Mamba architecture [117]. EHRMamba reduces the computational complexity with a more efficient attention scheme with linear complexity (instead of quadratic) and proposes a multi-task fine-tuning phase and a longer maximum sequence length of 2,048 tokens.

A significant challenge in this field is the lack of standardized benchmarks, which hinders fair model comparisons. Recently, van de Water et al. [335] introduced YAIB, a framework for reproducible cohort generation and data processing of public ICU datasets, along with a set of tasks for evaluating new models. Furthermore, EHRSHOT [344]

provides a dataset designed for model fine-tuning beyond ICU contexts.

ICU-BERT captures the complexity of health data in ICUs. It processes clinical variables with pre-trained input embeddings and models their values at the embedding level, incorporating temporal information. This approach is not limited to a narrow vocabulary of variables and is generalizable across different databases, as demonstrated in benchmarks against the YAIB framework and DuETT's tasks.

6.4.3 Methodology

ICU-BERT introduces a novel approach to capture the inherent complexity of medical data, intending to develop more robust DSSs that are generalizable across ICU and potentially broader medical contexts. The model is built upon the bidirectional transformer architecture introduced by BERT [76], optimized to handle the high-dimensional, sparse, and multivariate nature of ICU data, which is characterized by varying time resolutions and sampling rates.

We utilized several large, open-access ICU databases to train and validate ICU-BERT. The Medical Information Mart for Intensive Care (MIMIC)-IV database was employed during the pre-training phase, following a restructuring process designed to organize patient data more effectively for general medical applications. Subsequent fine-tuning was conducted using the MIMIC-III, eICU, and High Time Resolution ICU Dataset (HiRID) databases, applying an open-source pre-processing pipeline tailored to accommodate the specific requirements of each dataset.

A key innovation in ICU-BERT is the introduction of a multi-token input strategy, which encapsulates each medical record as a quadruplet comprising the clinical variable, its value, the time of recording, and the duration of the event. This approach enables ICU-BERT to capture detailed clinical information, ensuring a more refined understanding of patient data. An advanced embedding layer further processes this rich data structure, enhancing the model's capability to interpret complex clinical contexts.

In addition to its novel input strategy, ICU-BERT leverages pre-trained text embeddings from BioBERT [181], an LLM trained specifically on biomedical texts. These embeddings provide a robust initial representation of medical concepts, facilitating effective generalization across varied clinical environments and data formats. Traditional vocabularies often fail to capture the subtle nuances of clinical terms without extensive harmonization. For instance, variations in how blood pressure is measured, whether sitting, lying down or standing, are clinically relevant but might be misrepresented by conventional vocabularies.

ICU-BERT also incorporates a novel pre-training masking task and a multi-task loss to optimize learning its enriched multi-token representation. This approach allows for effective reconstruction of both clinical variables and their associated values during pre-training. The pre-training phase enables ICU-BERT to develop refined and adaptable data representations without requiring extensive pre-processing, such as semantic feature harmonization or feature engineering.

These innovations collectively enhance ICU-BERT’s ability to interpret the complex data landscapes typical of ICU environments, significantly advancing its utility for real-world clinical applications.

6.4.3.1 Datasets

With the increasing digitization of healthcare, ICUs have become data-intensive environments where continuous patient monitoring and automatic data collection are integral to care. Technologically advanced hospitals have adopted Health Information System (HIS) to systematically capture and manage this data, enabling the development of more robust DSSs. Several institutions have published anonymized ICU datasets to support research in this area, often hosted on platforms like PhysioNet [112].

Despite anonymization, medical data still carries significant privacy and security concerns. Access to these datasets is typically restricted, requiring individual approval for research use. Researchers must also complete training in human subject protection and data privacy and agree to the specific terms of the data license.

For the development of ICU-BERT, we utilized four major ICU datasets: MIMIC-IV, MIMIC-III, eICU, and HiRID. These databases provide a wide range of clinical data from different ICU environments and patient populations, allowing us to build and validate ICU-BERT across diverse settings.

In addition to the datasets used in this study, other valuable ICU datasets could enhance and generalize ICU-BERT in the future. For example, the AmsterdamUMCdb dataset [317] offers granular information from a single academic medical center in the Netherlands. The Pediatric Intensive Care (PIC) database [356], sourced from The Children’s Hospital at Zhejiang University School of Medicine, includes data from ICU stays between 2010 and 2018. The SICdb dataset [270], collected from University Hospital Salzburg in Austria, provides de-identified data, including monitor signals, lab results, and medication information. These datasets present opportunities for further research to improve the generalization and robustness of ICU-BERT across different ICU contexts.

MIMIC-IV

MIMIC-IV [151] is the latest iteration of the widely used MIMIC database, an evolution that incorporates contemporary data and introduces improvements over MIMIC-III. To develop ICU-BERT, we leveraged version 2.2, which contains data from 2008 to 2019. Recently, version 3.0 was launched with more data from 2020 to 2022 and some minor improvements. MIMIC-IV is sourced from two distinct in-hospital database systems at Beth Israel Deaconess Medical Center: a custom hospital-wide EHR system and an ICU-specific clinical information system. The database adopts a modular approach, making it easier to trace data provenance and facilitating the combination of disparate data sources:

- *hosp* Module: This module contains data from the hospital-wide EHR, predominantly collected during a patient’s hospital stay. It includes patient demographics (*patients* table), hospital admissions (*admissions*), intra-hospital transfers (*transfers*), laboratory measurements (*labevents*, *d_labitems*), microbiology cultures (*microbiologyevents*), medication prescriptions (*prescriptions*, *pharmacy*), and more. Timing information is de-identified and linked to a random anchor year between 2100 and 2200, with patient age provided relative to this anchor year. A real three-year group is provided to give some actual time information (e.g., 2008-2011). The maximum age is capped at 89 years, and death dates are recorded up to one year after hospital discharge.
- *icu* Module: This module is sourced from the ICU-specific clinical information system. Data are organized in a star schema, with *icustays* and *d_items* tables linking to event tables that document various clinical activities, such as intravenous inputs (*inputevents*), patient outputs (*outputevents*), procedures (*procedureevents*), and other charted information (*chartevents*).

MIMIC-IV is available in a character-based comma-delimited format, exported after the three following steps:

- Acquisition: Data were extracted for patients admitted to the emergency department or one of the ICUs. A master patient list contains all medical record numbers corresponding to these patients. The source tables were filtered to retain only data relevant to these patients.
- Preparation: The data were reorganized to facilitate retrospective analysis. This included de-normalizing tables, removing audit trails, and reorganizing into fewer tables. No specific data cleaning was performed to ensure the dataset reflects real-world clinical data as accurately as possible.
- De-identification: In compliance with the HIPAA, patient identifiers were removed and replaced with de-identified integer identifiers for patients, hospitalizations, and ICU stays. A free-text de-identification algorithm was also applied where necessary. To maintain temporal consistency within each patient’s records, dates and times were shifted randomly by a consistent number of days for each patient. To ensure privacy, the dates are not temporally comparable between different patients.

In addition to the raw data, we retrieved data from the MIMIC Code Repository [148], which provides algorithms to process and extract additional concepts from the data. This repository offers code for loading the data into a relational structure and extracting clinical concepts, including severity of illness scores, comorbidity status, administrative definitions of sepsis, organ failure scores, and treatment administration.

MIMIC-III

MIMIC-III v1.4 [150] is the predecessor to MIMIC-IV and contains data collected from patients admitted to the same hospital between 2001 and 2012. Although predating MIMIC-IV, MIMIC-III remains a valuable resource for clinical research and model development. Like MIMIC-IV, MIMIC-III was populated with data acquired during routine hospital care, ensuring no additional burden on caregivers or interference with their workflow.

Two critical care information systems were used during the data collection period: the Philips CareVue Clinical Information System and the iMDsoft MetaVision ICU. While most data was merged when building the MIMIC-III database tables, some data types differed significantly between the CareVue and MetaVision systems. As a result, these data were stored separately in tables.

MIMIC-III followed similar de-identification processes as MIMIC-IV to ensure patient privacy and compliance with the HIPAA. A rigorously evaluated de-identification system removed protected health information from free text fields, such as diagnostic reports and physician notes. This system relied on extensive dictionary look-ups and pattern-matching with regular expressions to accurately strip identifying details from the data.

eICU

The eICU Collaborative Research Database v2.0 [243] is a multi-center dataset containing de-identified health information from admissions to ICUs across the USA between 2014 and 2015. This extensive dataset encompasses vital sign measurements, care plan documentation, severity of illness measures, diagnostic information, and treatment details. The data is collected as part of the Philips eICU program, a critical care telehealth initiative that provides remote monitoring and decision support for bedside caregivers. While the program supplements on-site clinical teams, the data used by remote caregivers is archived for research purposes, providing a rich resource for investigating critical care practices.

The eICU database covers admissions to 335 units at 208 hospitals throughout the USA. The hospital from MIMIC is expressly excluded from the database. The dataset includes over 200,000 patient unit encounters for more than 139,000 unique patients. The data represents a broad range of hospital types and regions, making it one of the most diverse ICU datasets available.

Data within the eICU database includes a wide range of patient information, such as vital signs, laboratory measurements, medications, Acute Physiology and Chronic Health Evaluation (APACHE) scores, among others. Data from each patient is collected in a centralized warehouse if the appropriate interfaces are available within the hospital's EHR system. Since different care units may have varying levels of interface integration, some data may be unavailable for certain patients, even though those measurements might have been recorded in reality.

The eICU database is a relational database with multiple tables joined by key fields. Each table represents different aspects of patient care, and these tables can be linked using unique identifiers to reconstruct a comprehensive view of a patient’s ICU stay.

To comply with HIPAA, all personal health information has been removed from the dataset. Hospital and unit identifiers have also been anonymized to protect the privacy of participating organizations.

HiRID

HiRID v1.1.1 [95] is a publicly accessible critical care dataset that provides comprehensive information on nearly 34,000 patient admissions to the Department of Intensive Care Medicine at the Bern University Hospital in Switzerland. The hospital’s ICU, a 60-bed interdisciplinary unit, admits over 6,500 patients annually and offers a full range of modern intensive care services for adult patients. The dataset was developed in collaboration with the Swiss Federal Institute of Technology (ETH) Zürich, with the initial aim to support research into applications for early prediction of critical events such as circulatory failure [140].

HiRID has a high-resolution data collection, with measurements recorded every two minutes, which is significantly more frequent than other publicly available ICU datasets. The dataset contains 681 routinely collected physiological variables, diagnostic test results, and treatment parameters, covering January 2008 to June 2016. This high time resolution makes HiRID particularly valuable for studies focusing on continuous monitoring and early detection of clinical deterioration. The database offers two states of data: raw and pre-processed. Aside from necessary de-identification steps, the raw data is mainly untouched, providing researchers with 685 available variables. The data is organized into three tables: observations, pharmaceutical records, and general data.

Following other databases, HiRID also applies a strict anonymization procedure to ensure patient privacy. The dataset complies with the HIPAA requirements and the European Union’s GDPR standards, guaranteeing the de-identification of individuals in the dataset.

6.4.3.2 Data Processing

ICU Data Structuring

ICU-BERT introduces an advanced input data structure that enhances the processing of ICU data. By incorporating pre-trained embeddings, the model learns contextual representations from features and categorical values without being confined to a predetermined set of features.

The need to handle sparse and complex data prompted the development of a new structuring framework for MIMIC-IV, consolidating all data into a streamlined schema. While previous frameworks often overlooked the requirements for such generalization,

a recent initiative by researchers, the Medical Event Data Standard (MEDS) [18], aligns with the principles of ICU-BERT in addressing these challenges.

MIMIC-IV v2.2 organizes its relational database into two primary modules: *hosp* and *icu*. These modules differentiate based on the information system's variable types and sources. Each module consists of data tables formatted as character-based, comma-delimited files.

Every table in MIMIC-IV includes identifiers like *subject_id*, linking data to individual patients. Patients can have multiple hospital visits (*hadm_id*) and may be admitted to an ICU (*stay_id*) during each visit.

In the data schema proposed with ICU-BERT, each timestamped clinical registry is associated with the corresponding patient, providing a comprehensive dataset that includes static information (such as gender, as defined in the database) and detailed records of hospital and ICU stays. This chronological collection of data entries constitutes the patient's medical history. All data is retrieved from the database using either PostgreSQL or Google BigQuery queries, as detailed in the MIMIC-IV Code Repository [148]. Then, all resulting data frames are organized into a nested dictionary format for ease of access and manipulation, whose final structure is schematized in Listing 1.

Each medical registry is characterized based on its nature, whether isolated, such as a temperature measurement, or continuous, like intravenous medication. Registries are encapsulated in dictionaries that include the value, unit if applicable, and an extended description of the clinical variable when available. Registries are timestamped or associated with a tuple of timestamps, and all entries related to the same clinical variable are aggregated into a dictionary. An additional layer, the *source* level, organizes variables by original table and module, formatted as *module_table*. Both *source* and *variable* identifiers are integral in constructing the quadruplets used in ICU-BERT.

ICU-BERT extracts variables from multiple tables across both *hosp* and *icu* modules. From each table, relevant columns for registry extraction are identified. For instance, in the *admissions* table, the column names directly correspond to demographic variable keys, with each row providing the associated values. Conversely, in the *chartevents* table, each cell within the *itemid* column specifies a measurement converted into a variable key. Additionally, derived tables created by the MIMIC team, which contain developed medical concepts, are incorporated into the data framework (referenced in [148]). Table C.1 provides a comprehensive overview of all tables utilized by ICU-BERT.

Although this schema was initially designed for MIMIC-IV, it can be adapted in the future to other databases. Furthermore, it offers the flexibility to incorporate additional data fields, facilitating multimodal data integration in future work.

YAIB Processing

The proposed hierarchical schema in ICU-BERT enables the efficient structuring of all available variables for model training. However, adapting the original ICU databases to this standard structure involves a complex and time-consuming process. It requires

Listing 1: Definition of ICU-BERT data structure.

```
1 patients = {
2     patient_id : {
3         "general_info": {...},
4         "admissions": {
5             admission_id: {
6                 "general_info": {...},
7                 "transfers": {
8                     (admission time, discharge time): {
9                         "value": str,
10                    }
11                },
12                "services": {
13                    (admission, discharge): {
14                        "value": str,
15                        "description": str,
16                    }
17                },
18                "icu_stays_info": {
19                    icu_stay_id: {
20                        "admission_time": timestamp,
21                        "discharge_time": timestamp,
22                        "los": float,
23                    }
24                }
25            }
26        }
27        "medical_data": {
28            source: {
29                variable: {
30                    # Structure of a discrete event
31                    timestamp: {
32                        "value": str/float,
33                        "unit": str,
34                        "description": str,
35                    },
36                    # Structure of a continuous event
37                    (timestamp start, timestamp end): {
38                        "value": str/float,
39                        "unit": str,
40                        "description": str,
41                    }
42                }
43            }
44        }
45    }
46 }
```

extensive customization of database queries or data-loading techniques specific to each dataset’s structure, a task beyond the scope of this thesis.

To address the challenge of standardizing ICU data processing, van de Water et al. [335] introduced the YAIB framework. This Python library organizes data from multiple ICU databases into a unified tabular format, significantly streamlining the data preparation process. YAIB relies on the *ricu* R package [33] version 0.6.0, which provides predefined mappings for key variables across various ICU datasets. For instance, the administration of dextrose in MIMIC-IV is recorded using multiple *itemid* codes (220950, 228140, and 220952 in the *inputevents* table). YAIB consolidates these codes to capture all relevant values for each clinical variable.

ICU-BERT, in contrast, does not require such detailed mapping. By retrieving all unique *itemids* and leveraging pre-trained textual embeddings, the model learns the relationships between codes (e.g., associating all dextrose-related codes) based on their textual descriptions. This enables the model to handle a broader range of variables without manual aggregation or predefined mappings.

While the YAIB framework provides a harmonized structure for multiple ICU datasets, making data preparation more efficient, it does come with limitations. Specifically, it restricts the number of variables to those that can be consistently mapped across datasets, omitting many potentially useful but less standardized data points. Despite this, by enabling standardized comparisons, YAIB offers a valuable resource for validating ICU-BERT’s performance against more traditional pipelines.

Using the YAIB and *ricu* standards, we extracted harmonized datasets from MIMIC-IV (for comparison with our proposed method), MIMIC-III, eICU, and HiRID. These datasets include 52 features covering static and time series data such as vital signs, laboratory results, and input/output measurements, detailed in Table C.2.

6.4.3.3 Notation

In ICU-BERT, we conceptualize medical data for all P patients in a database. V_p integrates all ICU stays of patient $p \in \{1, 2, \dots, P\}$. Each ICU stay $v \in V_p$ contains a set R_v , which includes all registries r recorded during that stay. Each registry $r_i, i \in \{1, 2, \dots, N\}$, with N the total number of observations, is characterized by four attributes: the data source s , the clinical variable c , the recorded value x , and the associated timestamp or interval t . Mathematically, we can represent each registry as $r_i = (s_i, c_i, x_i, t_i)_{i=1}^N$.

To manage temporal dynamics and granular details, each stay in the ICU v is segmented into 24-hour windows with no overlap $w_j, j \in \{1, 2, \dots, K_v\}$, with K_v the total number of windows within v . Each window w_j includes all data registries r_i for which the timestamp t_i falls within the predefined start and end boundaries of the window, in addition to static demographic data, ward at the beginning, and previous diagnoses if documented.

Each registry r_i within a window w_j is transformed into a quadruplet representation defined as $q_i = (f(s, c), x, \tau(t), \delta(t))$, thus representing a token. $f(s, c)$ represents the

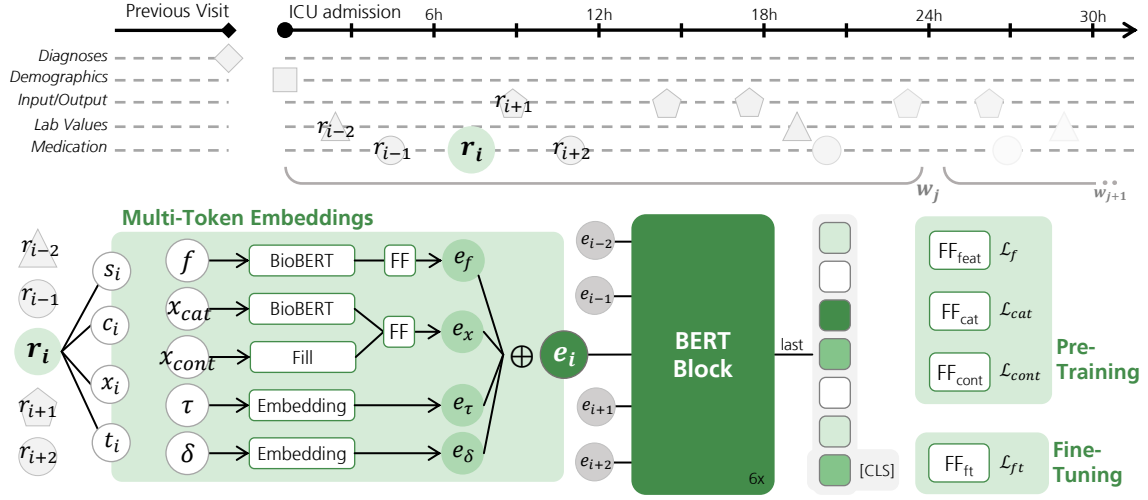


Figure 6.2: ICU-BERT architecture scheme. Complex, multivariate, and sparse medical registries r_i are processed by a multi-token embedding structure that combines embeddings from feature names f , categorical or numerical values x , timestamps τ , and durations δ . Pre-trained embeddings enhance the representations of features and categorical values, and a novel pre-training multi-task loss optimizes the simultaneous reconstruction of both features and values.

feature from the textual aggregation of source s and clinical variable c , $\tau(t)$ converts the timestamp t to encode the relative time since w_j start, and $\delta(t)$ calculates the duration associated with t , providing contextual timing information, assuming the value of 0 in case of a discrete event. The classification token, [CLS], and padding tokens, [PAD], are added, matching the quadruplet structure.

Figure 6.2 illustrates the architecture of ICU-BERT. Throughout a patient's stay in the ICU, numerous data points are recorded across various tables, as exemplified in the schematic. Each medical record r is organized into a sequence, capturing all parameters collected within the corresponding time window w_j . These records are then transformed into a structured quadruplet format before passing through the model's multi-token embedding layer.

6.4.3.4 Multi-Token Embeddings

To enhance the representational power of ICU-BERT and the ability to interpret and process medical terminology and context effectively, each $f(s, c)$ within the quadruplets q_i is enriched with BioBERT pre-trained embeddings, extracted from the token [CLS] from the last hidden state, such as:

$$PreE_f = \text{BioBERT}(f(s, c))[\text{CLS}] \quad (6.1)$$

For each quadruplet q_i , a Boolean mask m_i denotes whether the corresponding value x is continuous ($m_i = 1$). BioBERT embeddings are extracted for categorical values,

while a repetition of the continuous values matches the 768 dimension of the pre-trained embedding:

$$PreE_x = \begin{cases} \text{fill}(x), & \text{if } m_i = 1 \\ \text{BioBERT}(x)[\text{CLS}], & \text{if } m_i = 0 \end{cases} \quad (6.2)$$

Following the initial extraction of pre-trained embeddings, ICU-BERT employs a refined embedding layer to adapt these embeddings for ICU data analysis. Pre-trained embeddings of features f and values v undergo a dense transformation to d , the input dimension of the BERT block:

$$e_f = W_f \cdot PreE_f + b_f \quad (6.3)$$

$$e_x = W_x \cdot PreE_x + b_x \quad (6.4)$$

W_f and W_x represent the weight matrices for features and values, respectively, and b_f and b_x are the bias terms.

ICU-BERT employs an embedding layer that maps each discrete time-related input, in minutes, to a high-dimensional continuous vector for encoding temporal information such as timestamps and durations. Specifically, the embedding for time $\tau(t)$ and duration $\delta(t)$ are obtained by indexing into a pre-defined embedding matrix of input size 1,440 (minute-level in 24 hours), which transforms into d , represented as:

$$e_\tau = \text{Embedding}(\tau(t)) \quad (6.5)$$

$$e_\delta = \text{Embedding}(\delta(t)) \quad (6.6)$$

The embeddings e_f , e_x , e_τ and e_δ are summed to produce the composite embedding of r_i :

$$e_i = e_f + e_x + e_\tau + e_\delta \quad (6.7)$$

The resulting vector is subjected to dropout followed by layer normalization to ensure that the embeddings have a consistent scale and distribution. e_i thus encapsulates a detailed and nuanced representation of static and dynamic patient data, enabling the model to deliver precise and context-aware predictions in ICU settings.

6.4.3.5 ICU-BERT Configuration

ICU-BERT uses a BERT architecture, as described by [76] in their 2019 foundational paper. Our model adopts the original configuration of six transformer layers, each comprising multi-head self-attention mechanisms and fully connected networks. By applying this well-established architecture to ICU data, ICU-BERT benefits from BERT's powerful feature extraction capabilities, making it effective for interpreting complex medical data.

6.4.3.6 Pre-training Approach

Pre-training ICU-BERT is critical for adapting the generic language understanding capabilities of BERT to ICU contexts. ICU-BERT introduces a novel pre-training strategy, Masked Language-Value Modeling (MLVM). We follow BERT’s MLM by randomly selecting 15% of the quadruplets for masking. To enhance the model’s ability to reconstruct clinical features and their values, among the selected quadruplets, 50% have both their feature name and value masked, 25% have only the value masked, and 25% have only the feature name masked. Similar to MLM, selected features and/or values are masked with a [MASK] token, replaced by a random token from the relevant vocabulary, or left unchanged in the proportion 80%-10%-10%.

MLVM reconstruction is achieved through separate so-called vocabularies with feature names and values. The feature name vocabulary comprises all possible feature names, with size $\mathcal{F}$, while the value vocabulary includes all possible values of a given categorical feature, with size $\mathcal{V}$. Reconstruction is accomplished through three dedicated output heads. A feature classification head predicts the identity of masked features. It is equipped with an output dimension equal to the size of the feature name vocabulary to classify which feature was masked. The categorical value classification head is similar to the feature name head but predicts the identity of masked categorical values. It has an output dimension corresponding to the vocabulary size of the categorical values. Lastly, the continuous value regression head is distinctively designed to predict continuous values directly and outputs a single scalar.

Multi-Task Loss

ICU-BERT employs distinct loss functions in a multi-task framing to optimally train each classification head, reflecting the medical data’s varied nature. Cross-entropy loss is used to predict masked features and categorical values. However, the loss of categorical values must account for the continuous mask, m :

$$\mathcal{L}_f = -\frac{1}{K} \sum_{k=1}^K \sum_{c=1}^{\mathcal{F}} y_{k,c} \log(\hat{y}_{k,c}) \quad (6.8)$$

$$\mathcal{L}_{cat} = -\frac{1}{K - \sum_{k=1}^K m_k} \sum_{k=1}^K (1 - m_k) \sum_{c=1}^{\mathcal{V}} y_{k,c} \log(\hat{y}_{k,c}) \quad (6.9)$$

K is the total number of tokens in the batch. The reconstruction loss of continuous values uses the MAE:

$$\mathcal{L}_{cont} = -\frac{1}{\sum_{k=1}^K m_k} \sum_{k=1}^K m_k |y_k - \hat{y}_k| \quad (6.10)$$

Given the losses for features $\mathcal{L}_f$, categorical $\mathcal{L}_{cat}$, and continuous values $\mathcal{L}_{cont}$, the multi-task loss $\mathcal{L}_{total}$ becomes:

$$\mathcal{L}_{total} = \mathcal{L}_f + \beta \times \frac{\mathcal{L}_{cat} \times N_{cat} + \alpha \times \mathcal{L}_{cont} \times N_{cont}}{N_{cat} + N_{cont}} \quad (6.11)$$

α and β are parameterized weights to balance the contribution of value losses, while $N_{cont} = \sum_{k=1}^K m_k$ and $N_{cat} = K - N_{cont}$. This strategy not only challenges the model to predict the missing parts but also encourages it to learn robust associations between features and values.

6.4.3.7 Fine-tuning Approach

While the rich reconstructions from pre-training do not have direct clinical value, ICU-BERT can be fine-tuned to specific tasks. The original classification head needs to be replaced with an appropriate one and a suitable loss function. Outputs for fine-tuning are derived from the [CLS] token of the final transformer layer, which encapsulates the entire input sequence’s contextual information. This ensures that the predictions are based on comprehensive patient data, improving the applicability of ICU-BERT in clinical settings.

6.4.4 Experiments

Despite the limited availability of benchmarks for models like ICU-BERT [335], we conducted a set of experiments to assess its robustness and applicability.

We pre-trained ICU-BERT on MIMIC-IV after the data processing stage that structured the original relational database format into a hierarchical schema with timestamped entries for each clinical variable and data source, as detailed in Section 6.4.3.2. Using DuETT’s [174] data partition, we crafted distinct train, validation, and test sets in the proportion 70%-15%-15%.

For fine-tuning, ICU-BERT was adapted to a broad set of clinical tasks, including two from DuETT and three from YAIB [335], selected based on their relevance to demonstrate the versatility of the model. Moreover, to rigorously evaluate generalization capabilities, we employed a few-shot learning approach during fine-tuning. This approach was complemented by a zero-shot external evaluation and additional model fine-tuning across four additional datasets to assess generalization and the impact of feature reduction on performance. Using the previous partition, 5-fold CV was implemented by resampling train and validation sets.

6.4.4.1 Real-World Tasks

The fine-tuning process was carried out in five distinct classification and regression tasks to assess the predictive power of the model and its applicability to real-world clinical scenarios. From the four datasets processed with the YAIB framework, cohorts, and targets are predefined for each task, considering specific exclusion criteria. The same rationale was applied to fine-tuning samples for the proposed MIMIC-IV structuring schema.

Hospital Mortality

This task is based on the approach used in DuETT [174] and involves predicting the likelihood of a patient dying during their hospital stay based on the first 48 hours of data collected in the ICU. The task is structured as a binary classification problem, aiming to predict whether a patient will survive.

The authors of DuETT based this task on the work of Harutyunyan et al. [127] but adapted the exclusion criteria to include patients with multiple ICU stays or transfers between ICU units. This choice allows for a more comprehensive evaluation that reflects the real-world complexity of ICU care, where patients may be transferred between units based on clinical needs.

Phenotyping

Based on DuETT and the work of Harutyunyan et al. [127], this task is a multi-label classification problem that aims to identify which of 25 acute care conditions are present in a given patient's ICU stay record. Unlike traditional single-label classification, where only one outcome is predicted, this task requires simultaneously predicting the presence of multiple conditions. The task relies on the first 24 hours of an ICU stay, and the conditions are evaluated at the end of the patient's hospitalization.

The 25 phenotypes represent a range of acute care conditions that are critical in an ICU setting, such as acute respiratory failure, septic shock, and myocardial infarction. The complete list of phenotypes is available in Table C.3. Each phenotype is treated as a separate binary classification task, meaning the model must decide whether each condition is present or absent in the patient's record.

ICU Mortality

This task from YAIB focuses on predicting ICU mortality as a binary classification problem based on data collected within the first 24 hours of a patient's ICU stay. This task is particularly critical in real-time decision support, as early identification of at-risk patients allows clinicians to take preemptive actions that may reduce mortality.

Kidney Function

From YAIB, this regression task involves predicting the median creatinine levels on day two after ICU admission, based on the first 24 hours. This clinical parameter was selected as the most suitable proxy of kidney function.

AKI Onset

The onset of AKI is predicted using the Kidney Disease Improving Global Outcomes (KDIGO) 2012 criteria [158]. AKI is defined as KDIGO stage ≥ 1 , triggered either by

increased serum creatinine or low urine output. Although YAIB employs a many-to-many recurrent strategy that predicts the onset every hour based on all previous data, a continuous rolling window scheme is adopted in ICU-BERT, using a 24-hour window with a 6-hour step, to account for the maximum sequence length of 512 tokens.

Table 6.1 shows the summary of learning samples in each dataset from the cohort extraction for each task, displaying the high imbalance common in ICU-related tasks.

Table 6.1: Summary of total samples in included datasets and fine-tuning tasks. Classification tasks include the event’s positive rate (incidence), while the regression task (Kidney Function) includes the median value in mg/dL. The HiRID database does not publish patient identifiers.

	MIMIC-IV	MIMIC-IV	YAIB MIMIC-III	eICU	HiRID
Patients	50,920	50,920	46,476	160,816	N.A.
ICU Stays	73,181	73,181	61,532	182,774	32,338
ICU Mortality	49,833 (9.1%)	49,523 (7.2%)	38,128 (8.1%)	113,381 (5.5%)	12,859 (8.5%)
Kidney Function	33,677 (1.0)	33,949 (1.0)	27,481 (1.0)	69,116 (1.0)	7,499 (0.9)
Onset AKI	314,715 (13.1%)	478,582 (11.7%)	388,951 (8.7%)	1,177,049 (11.3%)	202,681 (7.8%)
Phenotyping	48,998 (18.6%)		N.A.		
Hospital Mortality	35,131 (14.2%)				

6.4.5 Implementation Details

ICU-BERT was implemented using PyTorch 2.3.0 alongside the HuggingFace Transformers library version 4.40.2 [343]. All development was performed on a Linux system utilizing CUDA 12.2 for GPU acceleration. Pre-training was conducted on a cluster equipped with eight NVIDIA Tesla A100 GPUs (each with 40GB of memory), using the SMX4 protocol and accessing 1.9 TB of VRAM in total. Fine-tuning was performed on a single-node NVIDIA Tesla A10 GPU setup, providing 205.4 GB of VRAM. A random seed of 42 was used across all experiments to ensure consistency in results.

ICU-BERT follows the BERT architecture with modifications inspired by the Med-BERT model [260]. The model consists of 6 encoder layers, each with a hidden size of 768, and 6 attention heads. The dense filter layer is set to a dimension of 64, and the model can handle sequences of up to 512 tokens.

6.4.5.1 Pre-training

Pre-training was conducted using the MIMIC-IV training dataset, with a vocabulary size $\mathcal{F} = 45,825$ for clinical features and $\mathcal{V} = 1,026$ for clinical values. The AdamW optimizer had a dropout rate of 0.1 and no weight decay. The learning rate was initialized at $5e-5$ and adjusted using a linear decay schedule, with a warmup phase spanning the first 40 epochs.

Hyperparameter tuning during pre-training was limited to a few key parameters due to the computational demands of training a large model. We followed the structural and parameter guidelines established by the original BERT [76] and Med-BERT [260] studies. Table C.4 lists the hyperparameters evaluated, the range tested, and the selected optimal values based on achieving the lowest multi-task loss (Equation 6.11) on the validation set. Default settings from the Transformers library’s BertConfig class were used for unspecified parameters. Pre-training parameter optimization utilized subsets of data comprising 10% of patients selected by their sequential patient identifiers. After initial tuning tests, the optimal hyperparameters α and β were defined as 3 and 1, respectively.

The model was trained over 102 epochs, requiring approximately 10.4 days. Due to budget constraints, the pre-training was manually halted before reaching the early stopping threshold.

6.4.5.2 Fine-tuning

After pre-training, ICU-BERT was fine-tuned to assess its performance over five different tasks. Employing a 5-fold CV setup, each task underwent fine-tuning five times per each few-shot percentage and dataset. Few-shot experiments were conducted on the MIMIC-IV dataset, using 1%, 2%, 4%, 8%, 16%, 32%, and 64% of the patient data. For each percentage level, the samples were randomly selected from the train set, and there is no guarantee that the samples in smaller subsets (e.g., 1%) are contained within the larger subsets (e.g., 2%), ensuring independence between different sampling percentages. A stratified split was applied to maintain target incidence for classification tasks, while no stratification was used for regression tasks.

Hyperparameter tuning was confined to the ICU mortality task during a 25-epoch run, using the Optuna 3.6.1 framework [11] for 200 iterations on a 10% patient sample, similar to the pre-training phase. This process also aimed to identify the optimal transformer layers for extracting representations and aggregating hidden states. The optimal settings were then applied to subsequent tasks. Table C.5 displays the tested hyperparameters and their optimal values, chosen to maximize the product of AUROC and AUPRC on the validation set.

The best results were achieved using the [CLS] token from the last transformer layer, with the final five layers of the model unfrozen during training. The AdamW optimizer was used for fine-tuning, with a learning rate of $1e-3$, a final head dropout rate of 0.5, and a linear learning rate scheduler without warmup.

6.4.6 Results

Following standard practices in the literature, classification task performance was evaluated using the AUROC and the AUPRC. For the regression task, MAE was used as the primary metric.

As depicted in Figure 6.3, ICU-BERT achieved an AUROC of $88.9 \pm 0.3\%$ and an AUPRC of $48.5 \pm 0.7\%$ for the ICU mortality prediction task when fine-tuned on the entire training dataset. The significant gap between these two metrics reflects the severe class imbalance inherent in ICU-related tasks, where the incidence rate is only 9.1%. Despite using class weighting to address this imbalance, the AUPRC remained disproportionately low, a trend observed across other classification tasks.

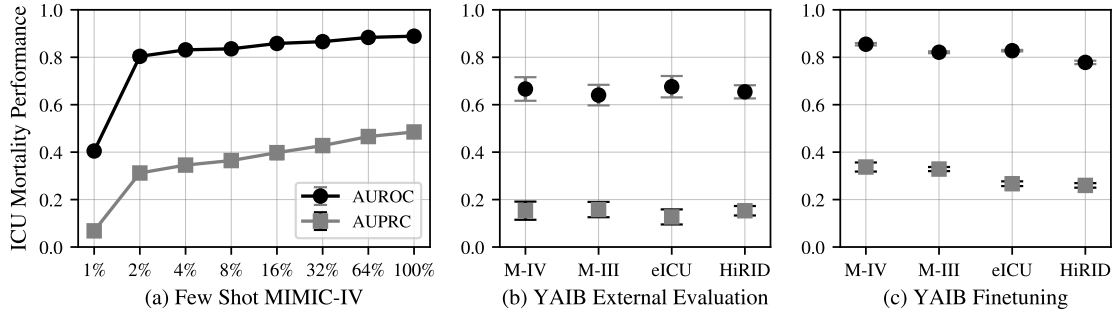


Figure 6.3: Results for the ICU mortality classification task in terms of AUROC and AUPRC (mean and standard deviation over 5-fold CV), on (a) Few-shot learning in the MIMIC-IV dataset using 1% to 100% of all patients in the train set; (b) External evaluation and (c) fine-tuning of the model trained with full MIMIC-IV train set. M-IV and M-III denote the YAIB MIMIC-IV and MIMIC-III datasets.

Regarding the few-shot experience, to avoid leakage, the few-shot sampling was done considering patient identifiers, which are proportional to the number of samples. For classification tasks, moving from 1% to 2% of training data significantly improves performance, with diminishing benefits as more data is added. Increasing the training set size from 2% to 100% for the ICU mortality task resulted in an improvement in AUROC of 8.6 percentage points and in AUPRC an increase of 17.3 percentage points. Although generally limited, there is a consistent improvement with each increment, highlighting the ongoing benefit of including more samples. This trend was also seen in the other tasks schematized in Figures 6.4 and 6.5, particularly in the continuous prediction of AKI onset. In this task, the AUROC improved from $66.0 \pm 0.7\%$ with only 1% of the data to $82.4 \pm 0.8\%$ with the entire dataset, while the AUPRC doubled. This more stable trend is likely due to the large volume of samples derived from each ICU stay (see Table 6.1). For the kidney function regression task, substantial MAE improvements were observed up to 16% of the training data, with diminishing returns after that (from 0.284 to 0.248 mg/dL).

Table 6.2 provides detailed results for both external evaluation and full-set fine-tuning in all YAIB datasets. It is important to note that the hospital mortality and phenotyping tasks, not included in YAIB’s original studies, were excluded from external validation in

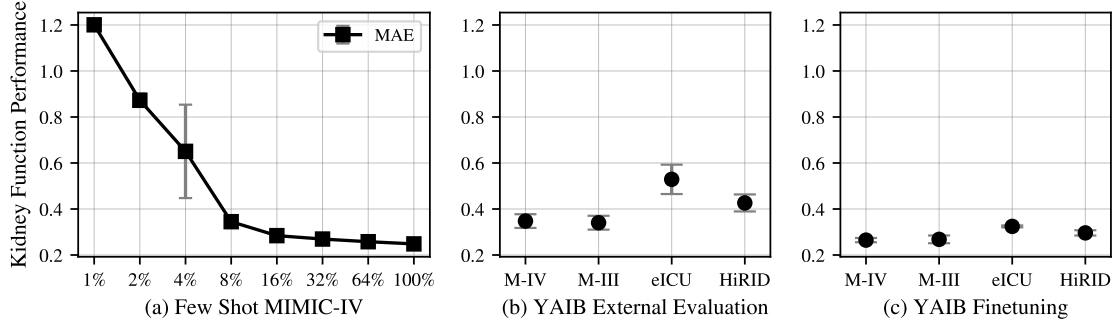


Figure 6.4: Results for the Kidney Function regression task in terms of MAE (mean and standard deviation over 5-fold CV), on (a) Few-shot learning in the MIMIC-IV dataset using 1% to 100% of all patients in the train set; (b) External evaluation and (c) fine-tuning of the model trained with full MIMIC-IV train set, in the YAIB datasets.

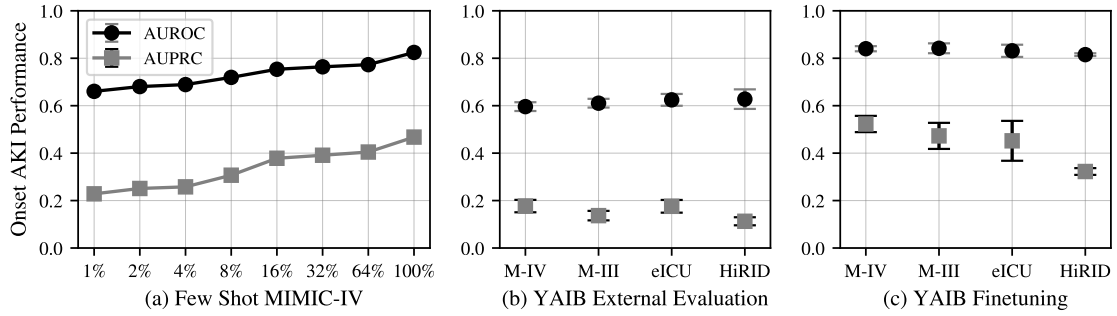


Figure 6.5: Results for the AKI Onset continuous classification task in terms of AUROC and AUPRC (mean and standard deviation over 5-fold CV), on (a) Few-shot learning in the MIMIC-IV dataset using 1% to 100% of all patients in the train set; (b) External evaluation and (c) fine-tuning of the model trained with full MIMIC-IV train set, in the YAIB datasets.

these datasets. The few-shot learning outcomes for both tasks are displayed in Figures C.1a and C.1b.

External evaluations using the YAIB datasets presented challenges for zero-shot generalization, with AUROC scores dropping to 64.1%-67.6% and AUPRC plummeting to 12.7%-15.8% in the ICU mortality prediction task. Similar performance declines were noted for AKI onset and kidney function tasks, especially when applying ICU-BERT to new datasets like eICU and HiRID. This drop reflects structural differences in dataset formats, distinct patient populations, and clinical settings, which ICU-BERT’s design mitigates by not being limited to a specific pre-training vocabulary. Other models would not be applicable unless extensive harmonization was conducted.

Despite these challenges, ICU-BERT demonstrated significant improvements when the output heads were fine-tuned specifically with the YAIB datasets, as shown in Figure 6.3(c). Performance improvements across all YAIB tasks and datasets confirmed ICU-BERT’s adaptability to various clinical scenarios, demonstrating progress toward a robust, foundational ICU model.

The success of ICU-BERT can be attributed to its multi-token embedding strategy,

Table 6.2: Performance of external evaluation and fine-tuning on the MIMIC-IV and YAIB datasets, across the YAIB tasks. Results are presented as the mean and standard deviation of AUROC and AUPRC, in percentage (%), for classification tasks, or MAE for the regression task. The best result for each task and dataset is emboldened.

Dataset	Mortality ICU		Onset AKI		Kidney Function
	AUROC	AUPRC	AUROC	AUPRC	MAE
ICU-BERT	88.9 ± 0.3	48.5 ± 0.7	78.0 ± 1.3	41.5 ± 2.1	0.248 ± 0.008
<i>YAIB External Evaluation</i>					
MIMIC-IV	66.6 ± 5.0	15.3 ± 3.8	59.6 ± 1.9	17.7 ± 2.6	0.348 ± 0.030
MIMIC-III	64.1 ± 4.3	15.8 ± 3.2	61.1 ± 1.8	13.7 ± 2.0	0.340 ± 0.030
eICU	67.6 ± 4.5	12.7 ± 3.2	62.5 ± 2.5	17.6 ± 2.7	0.529 ± 0.064
HiRID	65.4 ± 2.8	15.3 ± 2.0	62.8 ± 4.1	11.3 ± 1.7	0.426 ± 0.037
<i>YAIB Fine-Tuning</i>					
MIMIC-IV	85.5 ± 0.5	33.7 ± 1.9	84.0 ± 1.1	52.3 ± 3.4	0.265 ± 0.009
MIMIC-III	82.1 ± 0.4	32.9 ± 0.9	84.2 ± 2.1	47.3 ± 5.5	0.268 ± 0.017
eICU	82.8 ± 0.3	26.7 ± 1.0	83.1 ± 2.6	45.2 ± 8.4	0.324 ± 0.004
HiRID	65.4 ± 2.8	15.3 ± 2.0	62.8 ± 4.1	11.3 ± 1.7	0.426 ± 0.037

which captures the complexity of clinical variables at both semantic and numerical levels. By processing sparse sequential inputs directly from raw medical registries, ICU-BERT bypasses extensive data preprocessing and allows the model to learn from all available information, enhancing representation learning. Performance comparisons using two versions of MIMIC-IV further emphasized the importance of a rich feature set. Limiting the dataset to 52 features resulted in significant drops in ICU mortality and kidney function prediction accuracy (see Table 6.2).

Interestingly, fine-tuning on the AKI onset task using YAIB datasets led to better performance compared to MIMIC-IV, with improvements of up to 6.2% in AUROC and 10.7% in AUPRC. This could be due to the more targeted variable selection in the YAIB datasets, particularly for KDIGO-based AKI classification, which relies heavily on creatinine levels and urine output—features that are part of the smaller feature set. However, the utility of this task remains limited since KDIGO can be computed directly at the end of the observation window.

6.4.6.1 Literature Comparison

Comparing ICU-BERT with previous models presents challenges due to significant differences in application focus, variable counts, and window segmentation schemes used in ICU settings. In MIMIC-IV, we replicated the cohorts for each task described in van de Water et al. [335]. Table 6.3 summarizes ICU-BERT’s performance relative to baseline models from YAIB, including LR, elastic net, LightGBM [159], GRU [66], LSTM [133], Temporal Convolutional Network (TCN) [178], and a transformer model [326]. ICU-BERT

set new benchmarks for ICU mortality prediction and closely matched performance in the kidney function task. However, it did not outperform all models in the AKI onset prediction task, highlighting the challenges of adapting the BERT architecture for long input sequences, which are limited to 512 tokens after BERT’s input constraints.

Table 6.3: Performance of ICU-BERT in the YAIB tasks after fine-tuning on MIMIC-IV. The remaining models are baselines retrieved from [335]. Results are presented as the mean and standard deviation of AUROC and AUPRC, in percentage (%), for classification tasks, or MAE for the regression task. The best result is emboldened. N.A. indicates that the model was not trained in the task.

Model	Mortality ICU		Onset AKI		Kidney Function
	AUROC	AUPRC	AUROC	AUPRC	MAE
ICU-BERT	88.9 ± 0.3	48.5 ± 0.7	82.4 ± 0.8	46.8 ± 1.7	0.248 ± 0.008
LR	86.1 ± 0.1	39.7 ± 0.6	77.1 ± 0.2	37.7 ± 0.3	N.A.
LightGBM	87.7 ± 0.2	44.2 ± 0.7	83.8 ± 0.1	53.3 ± 0.2	0.24 ± 0.00
GRU	87.6 ± 0.1	42.8 ± 0.3	90.7 ± 0.1	69.6 ± 0.2	0.30 ± 0.01
LSTM	86.7 ± 0.4	41.0 ± 0.7	89.7 ± 0.1	66.5 ± 0.2	0.28 ± 0.01
TCN	87.1 ± 0.3	41.4 ± 0.8	89.8 ± 0.1	66.8 ± 0.2	0.28 ± 0.01
Transformer	86.9 ± 0.3	42.2 ± 0.3	89.6 ± 0.1	65.6 ± 0.2	0.32 ± 0.01
Elastic Net		N.A.			0.25 ± 0.00

To accommodate the maximum sequence size limitation, ICU-BERT was pre-trained using 24-hour windows. Although fine-tuning tasks focused on the initial 24 or 48 hours of ICU stay, the AKI onset task is continuous. While recurrent models, such as those in Table 6.3, excel in capturing temporal continuity through structured input matrices, they often fail to fully exploit the complexity of detailed patient data. With its advanced embedding strategy, ICU-BERT offers a promising alternative for tackling these challenges in clinical DSSs.

It is challenging to compare ICU-BERT with previous models due to the structural differences and experimental setups. Nevertheless, we recognize the importance of comparing the performance of new proposals to previous work. We tried to mirror the training partitions and classification tasks used in the DuETT framework with the MIMIC-IV dataset. Although there is no guarantee of duplicating their exact training process or using their precise task cohorts, we tried to fairly compare by replicating the conditions under which DuETT was tested while acknowledging the inherent limitations of such comparisons. The results, detailed in Table 6.4, show ICU-BERT surpassing other models in AUROC for the phenotyping task using only 24 hours of data, compared to full ICU stays in other models. However, its AUPRC scores were 3.1 percentage points lower than those reported for DuETT, with similar trends observed in the hospital mortality prediction metrics.

Table 6.4: Performance of ICU-BERT model in the tasks replicated from DuETT after fine-tuning on the MIMIC-IV. The remaining models are baselines published by [174]. Results are presented as the mean and standard deviation of AUROC and AUPRC, in percentage (%). The best result is emboldened.

Model	Hospital Mortality		Phenotyping	
	AUROC	AUPRC	AUROC	AUPRC
ICU-BERT	87.5 $\pm$ 0.7	55.8 $\pm$ 1.4	84.4 $\pm$ 0.1	57.3 $\pm$ 0.3
DuETT	91.2 $\pm$ 2.0	62.7 $\pm$ 0.2	83.8 $\pm$ 0.1	60.4 $\pm$ 0.3
STraTS [318]	88.2 $\pm$ 0.4	55.2 $\pm$ 1.3	82.0 $\pm$ 0.1	56.5 $\pm$ 0.6
Raindrop [361]	87.8 $\pm$ 0.1	54.6 $\pm$ 0.2	82.4 $\pm$ 0.1	57.7 $\pm$ 0.2
mTAND [293]	86.4 $\pm$ 0.2	54.0 $\pm$ 0.7	81.2 $\pm$ 0.1	55.3 $\pm$ 0.7
LSTM	88.1 $\pm$ 0.1	52.2 $\pm$ 0.6	75.6 $\pm$ 0.2	44.7 $\pm$ 0.2
XGBoost [63]	88.6 $\pm$ 0.2	59.3 $\pm$ 0.4	82.9 $\pm$ 0.1	58.9 $\pm$ 0.9

The differences in underlying model architectures, data handling, and potential variations in dataset pre-processing mean that our comparisons, though structured to be as fair as possible, cannot fully replicate the conditions or outcomes of the original DuETT study. This acknowledgment is crucial for contextualizing the results and understanding the specific capabilities and advantages of ICU-BERT to existing models.

6.4.7 Discussion

ICU-BERT advances representation learning for complex multivariate data, a crucial component for decision support in intensive care. This approach overcomes the limitations of previous work, which often relies on predefined vocabularies, heavily processed data structures, and a restricted set of variables. ICU-BERT efficiently processes highly granular and continuous data, offering substantial improvements in ICU DSSs.

However, building a truly foundational ICU model that is robust and generalizable remains a significant challenge. Such a technology should be capable of accessing and leveraging the entirety of a patient’s data accumulated over their lifetime. Future research could focus on integrating hierarchical strategies that extend the model’s capacity to handle longer sequences. This might be achieved through the addition of recurrent layers or by employing advanced transformer architectures that can aggregate data across multiple time windows, similar to models like Hi-BERT or DuETT. Additionally, incorporating newer architectures designed to handle longer sequences, such as Longformer [29] or Mamba [117], could further enhance ICU-BERT’s performance. These approaches could also help mitigate the challenges associated with segmenting continuous tasks into discrete windows, thereby addressing issues related to the increased number of samples, as seen in Table 6.1.

Comparing ICU-BERT to existing models reveals some inherent challenges. For instance, Med-BERT primarily focuses on processing ICD codes at the visit level, which

limits its application to long-term outcome prediction. In contrast, ICU-BERT is designed to handle the granular and continuously collected data typically found in ICU settings, making it well-suited for real-time alarm systems. When comparing ICU-BERT to DuETT, which also targets continuous ICU data, we replicated partitions and tasks for the MIMIC-IV dataset. However, replicating DuETT’s training conditions and patient cohorts was not entirely achievable. While ICU-BERT may not consistently outperform all existing models, its adaptability to new clinical contexts without requiring extensive preprocessing is a significant advantage. The model’s multi-token input strategy and pre-trained textual embeddings enable it to capture the semantic relationships between features and values, moving beyond the constraints of predefined vocabularies traditionally used in medical models.

Despite these advancements, building a foundational ICU model remains a complex challenge. Extensive hyperparameter tuning, which was not feasible in this study, could potentially enhance performance. For example, relying solely on the [CLS] token from the last layer might not fully capture the complexities of the sequence data, mainly since pre-training was conducted using the MLM task without incorporating tasks akin to NSP. Furthermore, while pre-trained embeddings improve the extraction of detailed information from features and help distinguish between continuous and categorical variables, they also increase input dimensions, impacting computational efficiency. Future work could explore pooling mechanisms or using smaller, more efficient pre-trained models. Continuous value embeddings might also be optimized using advanced strategies such as those proposed by Gorishniy et al. [115].

Initial tests suggest that unfreezing specific BERT layers during fine-tuning can enhance performance, although this may come with higher computational costs. While ICU-BERT effectively handles multivariate data, it has not yet fully integrated multimodal information, such as unstructured clinical notes. Future research should focus on extending pre-training with additional databases to enhance the model’s capabilities and robustness across different clinical scenarios.

6.5 Final Remarks

This chapter focused on a critical challenge in developing robust DSSs in healthcare: the longitudinal nature of medical data, which is often sparse, asynchronous, and irregular. While previous chapters briefly addressed this issue, the complexity and unique challenges associated with longitudinal data required a more in-depth exploration. The contributions of this chapter are summarized in Figure 6.6.

To address this, we proposed ICU-BERT, a proof of concept for adapting transformer architectures to handle the complexity of medical data. ICU-BERT represents a significant step toward building more robust DSSs by leveraging the full scope of available data and accounting for the intricate and varied nature of ICU data. By employing a multi-token input strategy and pre-trained textual embeddings, ICU-BERT captures detailed semantic

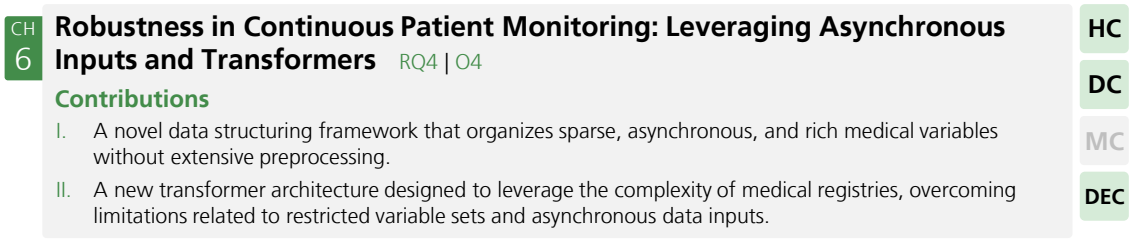


Figure 6.6: Overview of Chapter 6 contributions, aligned with the respective approaches of the holistic framework proposed in Chapter 3.

and numerical relationships across a diverse set of variables, improving the model’s ability to interpret longitudinal and complex medical data. Its performance in several tasks demonstrated the potential of this approach, although much work remains to be done.

There are several avenues for future research to achieve more robust medical foundational models, such as extending their capacity to handle longer windows of data or entire patient histories, incorporating multimodal data like clinical notes or imaging, and generalizing the model to broader populations and tasks. Moreover, addressing computational efficiency, explainability, and uncertainty estimation is essential for the future applicability of such models in real-world clinical settings.

The potential for such a general model is enormous. ICU-BERT takes the first steps to create a robust and adaptable system to virtually any real-world scenario, serving as a true digital twin of each patient by continuously learning from their entire medical history. However, such a powerful technology raises ethical questions and concerns about the trustworthiness of consequent DSSs. Generative models can create risks concerning patients’ privacy and safety without careful consideration. These issues are not trivial and must be carefully considered and addressed as the work progresses. In this thesis’s next and final chapter, we will reflect on these challenges and explore the broader implications of developing such advanced DSSs in healthcare.

CONCLUSION

Throughout this thesis, we addressed key challenges and opportunities in developing robust AI-based clinical DSSs for real-world healthcare applications. In response to the research questions posed in Chapter 1, we explored how the robustness of clinical DSSs can be enhanced by adopting a holistic approach, incorporating human-, data-, model-, and deployment-centric strategies. This framework guided our exploration across several medical use cases: risk estimation in cardiothoracic surgery, cardiovascular disease classification from ECG data, and continuous monitoring in the ICU. Each of these use cases introduced distinct technical challenges related to robustness, aligned with the objectives defined by our research questions.

RQ1. What are the key components of a comprehensive framework for developing robust AI in healthcare?

The potential of AI-based clinical DSSs spans across various domains, offering opportunities for more efficient administrative processes, improved early outcome prediction, personalized medicine, and enhanced patient engagement. However, significant challenges, including data quality, complexity, scalability, security, and user acceptance, continue to hinder the large-scale deployment of AI in healthcare. Ethical, social, and legal considerations further compound these challenges. In Chapter 3, we explored these opportunities and challenges, particularly focusing on the specificities of real-world medical data. We then proposed a comprehensive framework for developing robust AI-based clinical DSSs, emphasizing the importance of human-centric, data-centric, model-centric, and deployment-centric strategies. This holistic approach addresses the multi-faceted challenges in healthcare, ensuring that both technical and social aspects are considered. The framework presented in this chapter provides a foundation for tackling the specific challenges highlighted in the subsequent RQs.

RQ2. What are the optimal strategies to address the challenges of heterogeneous, complex, and incomplete real-world medical data in clinical DSSs?

Real-world medical data is often incomplete, fragmented, and inconsistent, making it difficult to train and deploy reliable AI-based clinical DSSs. Adequate data preparation is crucial for ensuring the robustness and reliability of these systems. In Chapter 4, we examined the challenges posed by real-world medical data, particularly in the field of cardiothoracic surgery, where extensive data from different stages of patient care is available but often incomplete or poorly structured. We collaborated with clinicians to design a clinician-in-the-loop framework for improving data preparation, which included strategies for handling missing data, imputing critical values, and structuring patient records for predictive modeling. Furthermore, we proposed new methods for missing data imputation based on correlation (CWKNNI, KNNSCI, and CWRI), in addition to applying similar strategies to longitudinal classification framed as a multi-label task with missing labels (CORKI). These efforts laid the groundwork for improving the data quality used in training and deploying robust clinical DSSs.

RQ3. How can clinical DSSs become more robust when dealing with multimodal data?

Healthcare data comes from various sources and can be interpreted in multiple representations, each requiring different approaches for integration and analysis. Many existing DSSs are designed to handle only one data type, limiting their effectiveness in real-world clinical settings. In Chapter 5, we investigated how clinical DSSs could be made more robust by incorporating multimodal data. Focusing on cardiovascular disease classification using ECG signals, we explored various fusion strategies to combine raw signals with signal-to-image transformations. By integrating multiple data representations, we investigated whether multimodal models can improve classification performance and contribute to the robustness of clinical DSSs. While multimodal fusion improved classification performance in some metrics, it did not lead to a significant overall improvement over the raw ECG models, highlighting the need for incorporating additional data modalities to achieve more robust models.

RQ4. Do inferences made by clinical DSSs improve when leveraging complex and asynchronous historical information?

In clinical settings, patient data is collected over time in a non-uniform and often asynchronous manner. This longitudinal data contains valuable insights into a patient's health trajectory, but effectively leveraging it for clinical decision-making remains a significant challenge, with previous work relying on restricted variables. In Chapter 6, we addressed the difficulties of handling sparse and asynchronous data in critical care, focusing on the ICU setting. We proposed ICU-BERT, a transformer-based architecture designed to process such data, incorporating historical patient information in its complete numerical

and semantical details to improve real-time decision-making. By adapting the input layer with a multi-token strategy and proposing an improved pre-training multitask loss, the model was able to provide good performance in distinct tasks in the context of continuous patient monitoring in the ICU. This work demonstrated the importance of incorporating complex and asynchronous temporal data to enhance the robustness and accuracy of clinical DSSs.

7.1 Limitations and Future Directions

Throughout this thesis, we have demonstrated how AI-based clinical DSSs can better manage the complexity of real-world healthcare data. Despite the contributions, several limitations persist that need further exploration in future research.

In Chapter 4, the clinician-in-the-loop strategy demonstrated the practical application of the proposed holistic framework for robust AI. However, the scalability of this approach needs further investigation, particularly regarding its economic feasibility. Highly pressured clinical environments may lack the resources to integrate local specificities into personalized DSSs. Nonetheless, a balance could be achieved, as certain characteristics are task-specific rather than context-dependent, suggesting that some standardization could reduce the need for extensive local adaptation.

On the other hand, the somewhat limited performance of the cardiothoracic surgery risk models of Chapter 4 suggests that certain clinically relevant variables may have been unavailable. These missing variables may be due to clinical parameters that are either not routinely recorded, reflecting limitations in current medical standards, or not integrated into existing medical information systems. While one could argue that more methodologies could have been explored, aligning the proposed model’s performance with results typically reported in the literature suggests more universal limitations. A more critical limitation emerges in Chapter 5, where our exploration of multimodal fusion in ECG classification was constrained by the nature of the data. Although we explored the combination of different representations of the same physiological data, additional modalities could have further demonstrated the potential of multimodal fusion. Unfortunately, such data is unavailable in current public ECG databases, aside from limited demographic information. In contrast, the ICU-BERT model presented in Chapter 6 integrated all available variables from ICU settings, but it remains limited to structured data types. Richer data sources were not incorporated, such as clinical notes, imaging, and physiological signals. The ongoing digitization of medical records, supported by initiatives such as the EHDS [91], may unlock new critical features, thus enabling further advancements in predictive accuracy.

General models for medical data hold tremendous potential for transforming clinical DSSs, as they could drastically reduce the effort required to develop specialized models for individual tasks. Since such a model would have knowledge of the interaction between

physiological parameters and pathophysiological mechanisms, using only a small set of labeled examples in a few-shot fine-tuning scheme could be sufficient for robust performance. Although ICU-BERT addresses some of the data compartmentalization challenges with a general data structure scheme proposal, future research must address technical limitations in handling multimodal sources and longer patient histories to process comprehensive time series data across a patient’s entire medical trajectory.

Regarding validation, the risk prediction DSS for cardiothoracic surgery was tested in a real-world setting, a significant milestone rarely achieved in the literature. However, this validation occurred in a single center, limiting the model’s applicability to the specific population of that hospital. Future work on the cardiothoracic surgery use case should extend to multi-center or multi-country settings to evaluate transferability toward building a more generalizable and effective DSSs [342]. While promising, our research in Chapters 5 and 6 remains exploratory and has not reached the stage of real-world deployment. The ECG classification and ICU monitoring use cases must incorporate expert feedback and undergo iterative validation in clinical environments to confirm their utility and practical effectiveness. Beyond real-world model validation, another future challenge is the long-term validity of models as clinical data evolves, medical practices change, and patient demographics shift. Future work should investigate continuous learning approaches that allow AI models to adapt over time without complete retraining.

While this thesis explores several strategies to enhance robustness, other mechanisms crucial for ensuring robust and trustworthy AI-based DSSs remain unexplored. Even in advanced *future* AI systems that possess superintelligence [45], models are susceptible to failure. Future work must integrate uncertainty estimation methods that allow users to recognize when predictions are unreliable or unsafe. Bias is also a critical concern, as it can cause models to perform poorly for underrepresented groups, exacerbating inequalities and undermining trust. Future research should include fairness audits, bias detection tools, and mitigation strategies. Additionally, in line with the ethical principle of explicability outlined by the European Commission’s Ethics Guidelines for Trustworthy AI [130], future DSSs should incorporate explanation mechanisms that provide transparency into the model’s decision-making process. This will allow stakeholders to identify potential hallucinations, biases, or errors, enhancing trust and ensuring accountability in clinical decision-making. While technical explainability is one facet, the usability of DSSs in clinical workflows also depends on the ability of healthcare professionals to trust and meaningfully engage with these systems. Integrating clinical feedback loops, intuitive user interfaces, and decision pathways that align with clinical reasoning will be necessary for future systems. Moreover, human-AI collaboration strategies need to be studied further to ensure DSSs empower, rather than overwhelm, clinicians. Furthermore, broader concerns related to the ethical and responsible development of AI, including protections against misuse, privacy assurances, and safeguarding human autonomy, must be incorporated into future frameworks for robust AI.

Lastly, societal and environmental concerns must also be considered. While clinical

DSSs have the potential to impact global health outcomes positively, the cost of developing general models often involves immense computational resources and energy consumption. The infrastructure required for pre-training these models demands significant non-renewable natural resources, such as critical minerals and rare earth elements, which pose environmental sustainability risks [69]. Furthermore, data labeling efforts often rely on underpaid labor from developing regions that may not benefit directly from these technological advancements [69], widening the economic divide between developed and developing nations. Although this thesis did not directly address energy efficiency beyond computational optimizations, it is an essential area for future research. Additionally, the data labels used in this research were either derived from clinical practice or inferred from the data, thus raising no immediate ethical concerns.

7.2 Contributions

Throughout this thesis, several contributions have been made to the scientific community, both in advancing knowledge in AI-based DSSs and collaborating with others through teaching and mentorship. Figure 7.1 aligns them with the respective chapter, RQ and objective.

As part of my academic journey, I was a teaching assistant in two courses within the Master's program in Biomedical Engineering at Nova School of Science and Technology: Machine Learning in Biomedical Engineering and Medical Image Processing. These experiences allowed me to learn and guide and support students in understanding key concepts in AI and medical data analysis.

Moreover, I had the privilege of mentoring and accompanying four master's students during their research projects: Bruno Ribeiro, Isabel Curioso, Luís Araújo, and Carolina Carvalho. Working with them provided a collaborative environment where I could share insights and contribute to their academic development while learning from their work.

The work and collaborations related to this thesis resulted in journal publications and conference proceedings. Some of these publications have been directly discussed in the chapters of this thesis, while others, though related, were not presented in detail.

7.2.1 Journal Papers

- J1 M. Barandas, D. Folgado, **R. Santos**, R. Simão, and H. Gamboa. "Uncertainty-based rejection in machine learning: Implications for model development and interpretability." *Electronics*, vol. 11, no. 3, p. 396, 2022. [26]
- J2 I. Curioso, **R. Santos**, B. Ribeiro, A. V. Carreiro, P. Coelho, J. Fragata, and H. Gamboa. "Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-based Imputation." *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 6, p. 101562, 2023. [72]

CH 3	
Foundations of Robust AI in Healthcare	
RQ	1. What are the key components of a comprehensive framework for developing robust AI in healthcare?
O	1. Analyze clinical DSSs limitations and design a holistic framework to improve robustness.
C	I. A holistic framework for robust AI in healthcare, integrating human-centric, data-centric, model-centric, and deployment-centric strategies.
P	J1 C5 C6
CH 4	
Robustness in Cardiothoracic Surgery Risk Estimation: Addressing Data Heterogeneity and Missing Inputs	
RQ	2. What are the optimal strategies to address the challenges of heterogeneous, complex, and incomplete real-world medical data in clinical DSSs?
O	2. Explore the potential of a clinician-in-the-loop approach to address the limitations of real-world heterogeneous medical data in cardiothoracic surgery risk prediction.
C	I. A clinician-in-the-loop pipeline for robust data processing.
	II. An optimization pipeline for training ML classification models development in complex scenarios.
	III. Deployment of a DSS in a clinical study to identify high-risk patients for remote monitoring.
	IV. Supervised the development of three novel correlation-based missing imputation mechanisms for addressing missing data in the absence of expert inputs.
	V. A correlation-based label imputation technique to address missing labels in multi-label classification.
P	J2 J4 C2 C3 C4
CH 5	
Robustness in ECG Diagnosis: Evaluating Multimodal Data Fusion Techniques	
RQ	3. How can clinical DSSs become more robust when dealing with multimodal data?
O	3. Study the potential of multimodal data representations for cardiovascular disease diagnosis.
C	I. A comprehensive comparison of multimodal fusion strategies, evaluating the effectiveness of combining raw ECG data with signal-to-image transformations for cardiovascular disease classification.
P	J3 J5
CH 6	
Robustness in Continuous Patient Monitoring: Leveraging Asynchronous Inputs and Transformers	
RQ	4. Do inferences made by clinical DSSs improve when leveraging complex and asynchronous historical information?
O	4. Incorporate rich, asynchronous data for robust continuous monitoring in critical care.
C	I. A novel data structuring framework that organizes sparse, asynchronous, and rich medical variables without extensive preprocessing.
	II. A new transformer architecture designed to leverage the complexity of medical registries, overcoming limitations related to restricted variable sets and asynchronous data inputs.
P	C1

Figure 7.1: Overview of this thesis contributions (C), aligned by chapter with the respective RQ and objective (O), in addition to all published works (P).

- J3 D. Folgado, M. Barandas, L. Famiglini, **R. Santos**, F. Cabitza, and H. Gamboa. “Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series.” *Information Fusion*, vol. 100, p. 101955, 2023. [103]
- J4 **R. Santos**, B. Ribeiro, I. Sousa, J. Santos, F. Guede-Fernández, P. Dias, A. V. Carreiro, H. Gamboa, P. Coelho, J. Fragata, and A. Londral “Predicting Post-Discharge Complications in Cardiothoracic Surgery: A Decision Support System to Optimize Remote Patient Monitoring Resources”. *International Journal of Medical Informatics*, vol. 182, p 105307, 2024. [280]
- J5 H. Narotamo[†], M. Dias[†], **R. Santos**[†], A. V. Carreiro, H. Gamboa, and M. Silveira. “Deep Learning for ECG Classification: A Comparative Study of 1D and 2D Representations and Multimodal Fusion Approaches”. *Biomedical Signal Processing and Control*, vol. 93, p. 106141, 2024. [227]

7.2.2 Conference Proceedings

- C1 B. Ribeiro, V. Cerqueira, **R. Santos**, and H. Gamboa. “Layered learning for acute hypotensive episode prediction in the ICU: an alternative approach.” *International Conference on e-Health and Bioengineering (EHB)*, IEEE, 2021. [266]
- C2 **R. Santos**, B. Ribeiro, P. Dias, I. Curioso, P. Madeira, F. Guede-Fernández, J. Santos, P. Coelho, I. Sousa, and A. Londral. “A Risk Prediction Framework to Optimize Remote Patient Monitoring Following Cardiothoracic Surgery.” *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) Demo Track*, pp. 366-371, 2023. [278]
- C3 **R. Santos**, B. Ribeiro, I. Curioso, M. Barandas, A. V. Carreiro, H. Gamboa, P. Coelho, J. Fragata, and I. Sousa. “CORKI: A Correlation-driven Imputation Method for Partial Annotation Scenarios in Multi-Label Clinical Problems”. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2023, PharML: Applications of Machine Learning in Pharma and Healthcare. Communications in Computer and Information Science*, vol 2136. [279]
- C4 B. Ribeiro, I. Curioso, **R. Santos**, F. Guede-Fernández, P. Coelho, J. Santos, J. Fragata, A. Londral, and I. Sousa. “Unravelling Heterogeneity: A Hybrid Machine Learning Approach to Predict Post-Discharge Complications in Cardiothoracic Surgery.” *EPIA Conference on Artificial Intelligence, Lecture Notes in Computer Science*, vol. 14116, 2023. [267]
- C5 B. Ribeiro, V. Rolla, and **R. Santos**. “INCOGNITUS: A Toolbox for Automated Clinical Notes Anonymization.” *17th Conference of the European Chapter of the*

[†]First authors.

Association for Computational Linguistics: System Demonstrations, pp. 187-194, 2023. [265]

- C6 I. Curioso, B. Ribeiro, P. Matias, **R. Santos**, J. Sousa, J. Ferreira, H. Gamboa, and D. Belo. "Pattern Recognition and Classification of Low-Intensity Emotions from Physiological Data." International Symposium on Computer-Based Medical Systems, pp. 299-304, 2023. [73]

7.3 Collaborations

Throughout this journey, I have had the privilege of collaborating with various institutions and professionals whose contributions and insights have significantly shaped my research. My work has been primarily supported and facilitated by Fraunhofer Portugal AICOS, my employer and primary affiliation, which provided essential resources and an enriching environment for my studies. Additionally, I have been affiliated with the Laboratory for Instrumentation, Biomedical Engineering, and Radiation Physics (LIBPhys) at Nova School of Science and Technology, the institution where I pursued my PhD.

Several funded projects have also been crucial in advancing this research. The Cardio-Follow.AI project, supported by the Portuguese Foundation for Science and Technology (ref. DSAIPA/AI/0094/2020), was instrumental in providing a real-world use case. This project allowed me to work closely with researchers from the VOH.CoLAB and the clinical and nursing teams at HSM. These collaborations gave me valuable insights into the practical challenges of deploying clinical DSSs in healthcare settings. Additionally, I had the opportunity to participate in the Center for Responsible AI, an initiative supported by the Plano de Recuperação e Resiliência (ref. 62_C645008882-00000055). This project brought together key stakeholders in the AI domain across Portugal, including research centers, start-ups, and companies. Interacting with these diverse actors gave me invaluable insights into the ethical, technical, and societal dimensions of AI.

Beyond these projects, I was privileged to undertake two research stays that profoundly influenced my work. My first stay was at Fraunhofer SCAI (Bonn, Germany), where I collaborated with Prof. Holger Fröhlich and his team. Prof. Fröhlich's guidance played an essential role in shaping my research direction on longitudinal modeling with asynchronous inputs, and the collaboration with other researchers at Fraunhofer SCAI further enriched my work. My second research stay at the Deep-REAL Lab at the University of Delaware (DE, USA) allowed me to collaborate with Prof. Xi Peng and his team. This experience exposed me to different research methodologies and perspectives, broadening my understanding of the global challenges and opportunities in AI research. These research stays were supported by Fraunhofer Portugal AICOS, with additional support from the Luso-American Development Foundation for the research at Deep-REAL.

These collaborations and exchanges have not only enhanced my research but also deepened my understanding of the multidisciplinary nature of developing robust AI in

healthcare. The support and insights from these institutions and individuals have been invaluable throughout my PhD journey.

7.4 Personal Reflections on AI and/or Tomorrow

“The truth is that we’re at a critical juncture in the history of our species, and if we don’t act soon, we could inhabit a world we don’t recognize anymore.”

— Al Gore, *Interview with Maranda Pleasant* (2013) [242]

Anyone could have hardly missed *An Inconvenient Truth*, as it became an award-winning documentary on the efforts of former Vice President of the USA, Al Gore, in raising concerns about global warming and the risk it poses to humanity. While some argue that his claims were excessive, we have seen temperatures rising, glaciers melting, and extreme events reoccurring. Al Gore’s mission is far from over, but I believe we may be approaching another critical juncture that will define the future of humanity.

Humans are living beings endowed by evolution with intelligence, creativity, and sentience (not a restricted list). Our ancestors mastered fire, unlocking a series of evolutionary milestones that brought us to where we are today. As I write this closing note, I feel both excitement and concern about the future, and the reader may have their sentiments and reflections after reading this thesis. Yuval Noah Harari defines the cognitive revolution as the beginning of humankind, followed by the agricultural revolution and, much later, the scientific revolution, which led us to become the most powerful species on Earth¹.

As humans built a global society, several critical junctures have challenged our future and disrupted established concepts. As posed by Daron Acemoglu and James A. Robinson, these historical landmarks have the power to change societies in different ways, as seen in events like the Black Death (pandemics), the Industrial Revolution (technological progress), or World War II (conflicts)². Regardless of the authors’ economic theory, the point is that the way nations react to these critical moments defines their future.

The recent developments in AI are leading humanity to a critical juncture, where societies must prepare for the potential adverse consequences while embracing the positive impacts of such a powerful technology. Throughout this thesis, we have discussed the opportunities of AI in healthcare by predicting distant diseases, applying targeted treatments with virtually 100% effectiveness, and solving the economic sustainability of national healthcare systems. Beyond healthcare, society will change if conceptual promises of AI in science fiction become reality.

However, what these creative sci-fi works also present are critical dilemmas, making us question whether that day should ever come. The stories are endless, featuring humanoid robots that emotionally connect with humans and misguide them, AI systems that revolt and trap humans in controlled simulations, intelligent systems that help defeat enemies,

¹in *Sapiens - A Brief History of Humankind*, 2011

²in *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*, 2012

or systems that know the future and use that power to control humanity. These scripts are fictional, but the topics they raise are not that far-fetched.

In healthcare, AI can be misused, even negligently. AI-supported genetic manipulation could lead to "manual" selection, flawed DSSs could harm or deny access to specific races or minorities, or biological weapons could hurt populations with particular characteristics. The consequences extend to other domains as well: AI could (can) spread misinformation and destabilize democratic regimes, control freedom of thought and speech, be used in war with large-scale impacts, or undermine critical thinking by shaping learning and free will. There's also the risk of nations losing sovereignty to unethical corporations, accelerating global warming, and eroding group-specific moral foundations that have guided humanity forward.

While previous generations faced similar concerns without such advanced technology, we now face the risk of equipping ourselves with tools that could scale negative consequences. I don't believe stopping progress is the solution, as our creative human minds are driven to explore the unknown. Blocking access to tools is difficult, and there are benefits yet to be unlocked from technological progress. What is critical is ensuring that the benefits significantly outweigh the drawbacks on a multidimensional scale, even if those dimensions are perhaps indeterminable.

Beyond the technical framework proposed in this thesis for robust AI, the solutions to bring universal benefit of AI are being discussed and must continue to evolve. Regulatory mechanisms are being established to block or reduce incentives for harmful usage. Developed technologies should be made available for widespread benefit by preventing monopolies. Society must be educated about the characteristics and limitations of AI tools. Research needs to be encouraged to study the environmental, economic, social, and ethical impact of hypothetical systems before deployment.

I believe the sci-fi future is still distant, but now is the time to develop the tools and set the boundaries that will allow us to reap the benefits of AI while avoiding the negative consequences of inaction. Humans must remain central to this equation, and for that, humanity must come together.

BIBLIOGRAPHY

- [1] 104th United States Congress a. *Health Insurance Portability and Accountability Act of 1996*. 1996. URL: <https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996> (visited on 2024-09-03) (cit. on p. 38).
- [2] H. Abdi and L. J. Williams. “Principal component analysis”. In: *WIREs Computational Statistics* 2.4 (2010-07). Publisher: Wiley, pp. 433–459. ISSN: 1939-5108, 1939-0068. DOI: [10.1002/wics.101](https://doi.org/10.1002/wics.101) (cit. on p. 23).
- [3] L. A. Abdulla and M. S. Al-Ani. “A review study for electrocardiogram signal classification”. In: *UHD Journal of Science and Technology* 4.1 (2020), pp. 103–117. DOI: [10.21928/uhdjst.v4n1y2020.pp103-117](https://doi.org/10.21928/uhdjst.v4n1y2020.pp103-117) (cit. on p. 111).
- [4] M. D. Abràmoff et al. “Considerations for addressing bias in artificial intelligence for health equity”. In: *npj Digital Medicine* 6.1 (2023-09-12). Publisher: Nature Publishing Group, pp. 1–7. ISSN: 2398-6352. DOI: [10.1038/s41746-023-00913-9](https://doi.org/10.1038/s41746-023-00913-9) (cit. on p. 38).
- [5] J. N. Acosta et al. “Multimodal biomedical AI”. In: *Nature Medicine* 28.9 (2022-09), pp. 1773–1784. ISSN: 1078-8956, 1546-170X. DOI: [10.1038/s41591-022-01981-2](https://doi.org/10.1038/s41591-022-01981-2) (cit. on pp. 40, 106).
- [6] N. Ad et al. “Comparison of EuroSCORE II, Original EuroSCORE, and The Society of Thoracic Surgeons Risk Score in Cardiac Surgery Patients”. In: *The Annals of Thoracic Surgery* 102.2 (2016-08), pp. 573–579. ISSN: 00034975. DOI: [10.1016/j.athoracsur.2016.01.105](https://doi.org/10.1016/j.athoracsur.2016.01.105) (cit. on pp. 54, 56).
- [7] J. Adler-Milstein et al. “Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis”. In: *NAM Perspectives* (2022). ISSN: 2578-6865. DOI: [10.31478/202209c](https://doi.org/10.31478/202209c) (cit. on pp. 31, 34, 35).
- [8] D. W. Aha, D. Kibler, and M. K. Albert. “Instance-based learning algorithms”. In: *Machine Learning* 6.1 (1991-01-01), pp. 37–66. ISSN: 1573-0565. DOI: [10.1007/BF00153759](https://doi.org/10.1007/BF00153759) (cit. on p. 12).

- [9] Z. Ahmad et al. "ECG Heartbeat Classification Using Multimodal Fusion". In: *IEEE Access* 9 (2021), pp. 100615–100626. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2021.3097614](https://doi.org/10.1109/ACCESS.2021.3097614) (cit. on pp. 107, 111, 113–115, 117, 118, 123, 127).
- [10] A. Akbarnejad and M. S. Baghshah. "An Efficient Semi-Supervised Multi-label Classifier Capable of Handling Missing Labels". In: *IEEE Transactions on Knowledge and Data Engineering* 31 (2019), pp. 229–242. DOI: [10.1109/TKDE.2018.2833850](https://doi.org/10.1109/TKDE.2018.2833850) (cit. on p. 90).
- [11] T. Akiba et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019-07-25. DOI: [10.48550/arXiv.1907.10902](https://doi.org/10.48550/arXiv.1907.10902). arXiv: [1907.10902\[cs, stat\]](https://arxiv.org/abs/1907.10902) (cit. on p. 151).
- [12] H. Akoglu. "User's guide to correlation coefficients". In: *Turkish Journal of Emergency Medicine* 18.3 (2018-09), pp. 91–93. ISSN: 24522473. DOI: [10.1016/j.tjem.2018.08.001](https://doi.org/10.1016/j.tjem.2018.08.001) (cit. on p. 20).
- [13] E. A. P. Alday et al. "Classification of 12-lead ECGs: the Physionet/Computing in Cardiology Challenge 2020". In: *Physiological measurement* 41.12 (2020). Publisher: IOP Publishing, p. 124003. DOI: [10.1088/1361-6579/abc960](https://doi.org/10.1088/1361-6579/abc960) (cit. on p. 92).
- [14] J. Allyn et al. "A Comparison of a Machine Learning Model with EuroSCORE II in Predicting Mortality after Elective Cardiac Surgery: A Decision Curve Analysis". In: *PLOS ONE* 12.1 (2017-01-06). Ed. by A. Parolari. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0169772](https://doi.org/10.1371/journal.pone.0169772) (cit. on p. 56).
- [15] A. Altmann et al. "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10 (2010-05-15). Publisher: Oxford University Press (OUP), pp. 1340–1347. ISSN: 1367-4811, 1367-4803. DOI: [10.1093/bioinformatics/btq134](https://doi.org/10.1093/bioinformatics/btq134) (cit. on p. 23).
- [16] L. Alzubaidi et al. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions". In: *Journal of Big Data* 8.1 (2021-03-31), p. 53. ISSN: 2196-1115. DOI: [10.1186/s40537-021-00444-8](https://doi.org/10.1186/s40537-021-00444-8) (cit. on p. 15).
- [17] F. Amrollahi et al. "Leveraging clinical data across healthcare institutions for continual learning of predictive risk models". In: *Scientific Reports* 12.1 (2022-05-19), p. 8380. ISSN: 2045-2322. DOI: [10.1038/s41598-022-12497-7](https://doi.org/10.1038/s41598-022-12497-7) (cit. on pp. 26, 41).
- [18] B. Arnrich et al. "Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health". In: *ICLR 2024 Workshop on Learning from Time Series For Health*. 2024-03-08. URL: <https://openreview.net/forum?id=IsHy2ebjIG> (visited on 2024-08-09) (cit. on p. 142).
- [19] P. C. Austin. "A Tutorial on Multilevel Survival Analysis: Methods, Models and Applications". In: *International Statistical Review* 85.2 (2017), pp. 185–203. ISSN: 1751-5823. DOI: [10.1111/insr.12214](https://doi.org/10.1111/insr.12214) (cit. on p. 131).

- [20] D. Bahdanau, K. Cho, and Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016-05-19. arXiv: [1409.0473\[cs, stat\]](https://arxiv.org/abs/1409.0473). URL: <http://arxiv.org/abs/1409.0473> (visited on 2024-09-29) (cit. on p. 118).
- [21] I. Bala et al. “Ensuring Security and Privacy in Healthcare Systems: A Review Exploring Challenges, Solutions, Future Trends, and the Practical Applications of Artificial Intelligence”. In: *Jordan Medical Journal* 58.2 (2024-07-15). Number: 2. ISSN: 2664-8091. URL: <https://jjournals.ju.edu.jo/index.php/JMJ/article/view/2527> (visited on 2024-09-29) (cit. on p. 41).
- [22] L. Ball, F. Costantino, and P. Pelosi. “Postoperative complications of patients undergoing cardiac surgery”. In: *Current Opinion in Critical Care* 22.4 (2016-08), p. 386. ISSN: 1070-5295. DOI: [10.1097/MCC.0000000000000319](https://doi.org/10.1097/MCC.0000000000000319) (cit. on p. 48).
- [23] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019-02). Place: USA Publisher: IEEE Computer Society, pp. 423–443. ISSN: 0162-8828. DOI: [10.1109/TPAMI.2018.2798607](https://doi.org/10.1109/TPAMI.2018.2798607) (cit. on pp. 107–109, 111).
- [24] M. Barandas et al. “Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram”. In: *Information Fusion* 101 (2024-01-01), p. 101978. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2023.101978](https://doi.org/10.1016/j.inffus.2023.101978). (Visited on 2024-05-09) (cit. on pp. 8, 128).
- [25] M. Barandas et al. “TSFEL: Time Series Feature Extraction Library”. In: *SoftwareX* 11 (2020-01-01), p. 100456. ISSN: 2352-7110. DOI: [10.1016/j.softx.2020.100456](https://doi.org/10.1016/j.softx.2020.100456) (cit. on pp. 22, 106).
- [26] M. Barandas et al. “Uncertainty-Based Rejection in Machine Learning: Implications for Model Development and Interpretability”. In: *Electronics* 11.3 (2022-01-28), p. 396. ISSN: 2079-9292. DOI: [10.3390/electronics11030396](https://doi.org/10.3390/electronics11030396) (cit. on p. 163).
- [27] M. Barrett et al. “Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care”. In: *EPMA Journal* 10.4 (2019-12-01), pp. 445–464. ISSN: 1878-5085. DOI: [10.1007/s13167-019-00188-9](https://doi.org/10.1007/s13167-019-00188-9) (cit. on p. 33).
- [28] J. A. Behar, J. Levy, and L. A. Celi. *Generalization in medical AI: a perspective on developing scalable models*. 2023-11-09. DOI: [10.48550/arXiv.2311.05418](https://doi.org/10.48550/arXiv.2311.05418). arXiv: [2311.05418\[cs\]](https://arxiv.org/abs/2311.05418) (cit. on p. 32).
- [29] I. Beltagy, M. E. Peters, and A. Cohan. *Longformer: The Long-Document Transformer*. 2020-12-02. DOI: [10.48550/arXiv.2004.05150](https://doi.org/10.48550/arXiv.2004.05150). arXiv: [2004.05150\[cs\]](https://arxiv.org/abs/2004.05150) (cit. on p. 156).

- [30] E. Ben-Baruch et al. “Multi-label Classification with Partial Annotations using Class-aware Selective Loss”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022-06, pp. 4754–4762. ISBN: 978-1-66546-946-3. DOI: [10.1109/CVPR52688.2022.00472](https://doi.org/10.1109/CVPR52688.2022.00472) (cit. on pp. 89, 90, 102).
- [31] U. Benedetto et al. “Can machine learning improve mortality prediction following cardiac surgery?” In: *European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery* 58.6 (2020-12-01), pp. 1130–1136. ISSN: 1873-734X. DOI: [10.1093/ejcts/ezaa229](https://doi.org/10.1093/ejcts/ezaa229) (cit. on p. 57).
- [32] Y. Bengio, A. Courville, and P. Vincent. “Representation Learning: A Review and New Perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013-08). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1798–1828. ISSN: 1939-3539. DOI: [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50) (cit. on p. 133).
- [33] N. Bennett et al. “ricu: R’s interface to intensive care data”. In: *GigaScience* 12 (2022-12-28), giad041. ISSN: 2047-217X. DOI: [10.1093/gigascience/giad041](https://doi.org/10.1093/gigascience/giad041) (cit. on p. 144).
- [34] N. Bento et al. “Comparing Handcrafted Features and Deep Neural Representations for Domain Generalization in Human Activity Recognition”. In: *Sensors* 22.19 (2022-09-27), p. 7324. ISSN: 1424-8220. DOI: [10.3390/s22197324](https://doi.org/10.3390/s22197324) (cit. on p. 106).
- [35] J. Bergstra and Y. Bengio. “Random search for hyper-parameter optimization”. In: *J. Mach. Learn. Res.* 13 (2012), pp. 281–305. ISSN: 1532-4435. URL: <https://dl.acm.org/doi/10.5555/2188385.2188395> (visited on 2024-09-29) (cit. on p. 24).
- [36] E. S. Berner, ed. *Clinical decision support systems: theory and practice*. 2nd ed. Health informatics. New York, NY: Springer, 2007. 269 pp. ISBN: 978-0-387-33914-6 (cit. on p. 29).
- [37] E. S. Berner and T. J. La Lande. “Overview of Clinical Decision Support Systems”. In: *Clinical Decision Support Systems: Theory and Practice*. Ed. by E. S. Berner. New York, NY: Springer, 2007, pp. 3–22. ISBN: 978-0-387-38319-4. DOI: [10.1007/978-0-387-38319-4_1](https://doi.org/10.1007/978-0-387-38319-4_1) (cit. on p. 29).
- [38] D. Berrar. *Cross-validation*. 2024. URL: http://berrar.com/resources/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf (visited on 2024-09-17) (cit. on p. 23).
- [39] H. V. Bhagat and M. Singh. “NMVI: A data-splitting based imputation technique for distinct types of missing data”. In: *Chemometrics and Intelligent Laboratory Systems* 223 (2022-04), p. 104518. ISSN: 01697439. DOI: [10.1016/j.chemolab.2022.104518](https://doi.org/10.1016/j.chemolab.2022.104518) (cit. on pp. 72, 80).

- [40] A. Bhatti et al. "Attentive Cross-modal Connections for Deep Multimodal Wearable-based Emotion Recognition". In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. Nara, Japan: IEEE, 2021-09-28, pp. 01–05. ISBN: 978-1-66540-021-3. DOI: [10.1109/ACIIW52867.2021.9666360](https://doi.org/10.1109/ACIIW52867.2021.9666360) (cit. on p. 114).
- [41] D. S. Bitterman, H. J. W. L. Aerts, and R. H. Mak. "Approaching autonomy in medical artificial intelligence". In: *The Lancet Digital Health* 2.9 (2020-09-01), e447–e449. ISSN: 2589-7500. DOI: [10.1016/S2589-7500\(20\)30187-4](https://doi.org/10.1016/S2589-7500(20)30187-4) (cit. on p. 34).
- [42] J. C. Bjerring and J. Busch. "Artificial Intelligence and Patient-Centered Decision-Making". In: *Philosophy & Technology* 34.2 (2021-06-01), pp. 349–371. ISSN: 2210-5441. DOI: [10.1007/s13347-019-00391-6](https://doi.org/10.1007/s13347-019-00391-6) (cit. on p. 39).
- [43] K. Blagec et al. "Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals". In: *Journal of Biomedical Informatics* 137 (2023-01-01), p. 104274. ISSN: 1532-0464. DOI: [10.1016/j.jbi.2022.104274](https://doi.org/10.1016/j.jbi.2022.104274) (cit. on p. 3).
- [44] D. Borra et al. "On the Application of Convolutional Neural Networks for 12-lead ECG Multi-label Classification using Datasets from Multiple Centers". In: *2020 Computing in Cardiology Conference*. 2020-12-30. DOI: [10.22489/CinC.2020.349](https://doi.org/10.22489/CinC.2020.349) (cit. on p. 116).
- [45] N. Bostrom. *Superintelligence: paths, dangers, strategies*. Reprinted with corrections 2017. Oxford, United Kingdom: Oxford University Press, 2017. 328 pp. ISBN: 978-0-19-967811-2 (cit. on pp. 45, 162).
- [46] P. Bota et al. "Emotion Assessment Using Feature Fusion and Decision Fusion Classification Based on Physiological Data: Are We There Yet?" In: *Sensors* 20.17 (2020-08-21), p. 4723. ISSN: 1424-8220. DOI: [10.3390/s20174723](https://doi.org/10.3390/s20174723) (cit. on p. 114).
- [47] T. J. Bradshaw et al. "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging". In: *Radiology: Artificial Intelligence* 5.4 (2023-07). Publisher: Radiological Society of North America, e220232. DOI: [10.1148/ryai.220232](https://doi.org/10.1148/ryai.220232) (cit. on p. 23).
- [48] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001-10-01), pp. 5–32. ISSN: 1573-0565. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324) (cit. on p. 12).
- [49] M. Brundage. "Artificial Intelligence and Responsible Innovation". In: *Fundamental Issues of Artificial Intelligence*. Vol. 376. Series Title: Synthese Library. Cham: Springer International Publishing, 2016, pp. 543–554. DOI: [10.1007/978-3-319-26485-1_32](https://doi.org/10.1007/978-3-319-26485-1_32) (cit. on p. 41).

- [50] K. Bruynseels, F. Santoni de Sio, and J. van den Hoven. “Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm”. In: *Frontiers in Genetics* 9 (2018-02-13). Publisher: Frontiers. ISSN: 1664-8021. DOI: [10.3389/fgene.2018.00031](https://doi.org/10.3389/fgene.2018.00031) (cit. on p. 33).
- [51] N. T. Bui and G. S. Byun. “The comparison features of ECG signal with different sampling frequencies and filter methods for real-time measurement”. In: *Symmetry* 13.8 (2021-08). Publisher: MDPI AG. ISSN: 20738994. DOI: [10.3390/sym13081461](https://doi.org/10.3390/sym13081461) (cit. on p. 116).
- [52] S. V. Buuren and K. Groothuis-Oudshoorn. “mice : Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (2011). ISSN: 1548-7660. DOI: [10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03) (cit. on pp. 71, 80).
- [53] F. Cabitza and A. Campagner. “The need to separate the wheat from the chaff in medical informatics”. In: *International Journal of Medical Informatics* 153 (2021-09), p. 104510. ISSN: 13865056. DOI: [10.1016/j.ijmedinf.2021.104510](https://doi.org/10.1016/j.ijmedinf.2021.104510) (cit. on pp. 28, 55).
- [54] F. Cabitza, A. Campagner, and C. Simone. “The need to move away from agential-AI: Empirical investigations, useful concepts and open issues”. In: *International Journal of Human-Computer Studies* 155 (2021-11), p. 102696. ISSN: 10715819. DOI: [10.1016/j.ijhcs.2021.102696](https://doi.org/10.1016/j.ijhcs.2021.102696) (cit. on p. 32).
- [55] L. Cai, J. Gao, and D. Zhao. “A review of the application of deep learning in medical image classification and segmentation”. In: *Annals of translational medicine* 8.11 (2020). Publisher: AME Publications. DOI: [10.21037/atm.2020.02.44](https://doi.org/10.21037/atm.2020.02.44) (cit. on pp. 113, 117).
- [56] M. Carabantes. “Black-box artificial intelligence: an epistemological and critical analysis”. In: *AI & SOCIETY* 35.2 (2020-06-01), pp. 309–317. ISSN: 1435-5655. DOI: [10.1007/s00146-019-00888-w](https://doi.org/10.1007/s00146-019-00888-w) (cit. on p. 32).
- [57] L. A. Carrasco-Ribelles et al. “Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review”. In: *Journal of the American Medical Informatics Association* 30.12 (2023-12-01), pp. 2072–2082. ISSN: 1527-974X. DOI: [10.1093/jamia/ocad168](https://doi.org/10.1093/jamia/ocad168) (cit. on p. 40).
- [58] E. Caruso and A. Rossi Zadra. “The trade-off between costs and outcome after cardiac surgery. Evidence from an Italian administrative registry”. In: *Health Policy* 124.12 (2020-12), pp. 1345–1353. ISSN: 01688510. DOI: [10.1016/j.healthpol.2020.09.005](https://doi.org/10.1016/j.healthpol.2020.09.005) (cit. on pp. 48, 54).
- [59] A. Cascarano et al. “Machine and deep learning for longitudinal biomedical data: a review of methods and applications”. In: *Artificial Intelligence Review* 56.2 (2023-11-01), pp. 1711–1771. ISSN: 1573-7462. DOI: [10.1007/s10462-023-10561-w](https://doi.org/10.1007/s10462-023-10561-w) (cit. on pp. 3, 36, 37, 129–131).

- [60] N. V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002-06-01), pp. 321–357. ISSN: 1076-9757. DOI: [10.1613/jair.953](https://doi.org/10.1613/jair.953). arXiv: [1106.1813\[cs\]](https://arxiv.org/abs/1106.1813) (cit. on p. 22).
- [61] D. Che et al. “Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics”. In: *Software Tools and Algorithms for Biological Systems*. Vol. 696. New York, NY: Springer New York, 2011, pp. 191–199. DOI: [10.1007/978-1-4419-7046-6_19](https://doi.org/10.1007/978-1-4419-7046-6_19) (cit. on p. 12).
- [62] Z. Che et al. “Recurrent Neural Networks for Multivariate Time Series with Missing Values”. In: *Scientific Reports* 8.1 (2018-04-17), p. 6085. ISSN: 2045-2322. DOI: [10.1038/s41598-018-24271-9](https://doi.org/10.1038/s41598-018-24271-9) (cit. on p. 136).
- [63] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016-08-13, pp. 785–794. DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785). arXiv: [1603.02754\[cs\]](https://arxiv.org/abs/1603.02754) (cit. on p. 156).
- [64] Z. Cheng and Z. Zeng. “Joint Label-specific Features and Label Correlation for Multi-label Learning with Missing Label”. In: *Applied Intelligence* 50 (2020). Publisher: Springer, pp. 4029–4049. DOI: [10.1007/s10489-020-01715-2](https://doi.org/10.1007/s10489-020-01715-2) (cit. on p. 89).
- [65] Chief Scientist and Science Division (SCI) and Health Ethics & Governance (HEG). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. 2024. URL: <https://iris.who.int/bitstream/handle/10665/375579/9/789240084759-eng.pdf?sequence=1> (visited on 2024-09-03) (cit. on p. 43).
- [66] K. Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014-09-02. DOI: [10.48550/arXiv.1406.1078](https://doi.org/10.48550/arXiv.1406.1078). arXiv: [1406.1078\[cs, stat\]](https://arxiv.org/abs/1406.1078) (cit. on pp. 112, 154).
- [67] K. Cho et al. “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches”. In: *Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (2014-09). ISBN: 9781937284961 Publisher: Association for Computational Linguistics (ACL), pp. 103–111. DOI: [10.48550/arxiv.1409.1259](https://doi.org/10.48550/arxiv.1409.1259) (cit. on pp. 15, 112, 131).
- [68] P. N. M. P. Coelho et al. “Quality of life after elective cardiac surgery in elderly patients”. In: *Interactive CardioVascular and Thoracic Surgery* 28.2 (2019-02-01), pp. 199–205. ISSN: 1569-9285. DOI: [10.1093/icvts/ivy235](https://doi.org/10.1093/icvts/ivy235) (cit. on pp. 48, 50, 88).
- [69] K. Crawford. *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press, 2021. 327 pp. ISBN: 978-0-300-20957-0 (cit. on p. 163).

- [70] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. 1st ed. Cambridge University Press, 2000-03-23. DOI: [10.1017/CB09780511801389](https://doi.org/10.1017/CB09780511801389) (cit. on p. 13).
- [71] I. Curioso. “Delivering Reliable AI to Clinical Contexts: Addressing the Challenge of Missing Data”. Master Thesis. NOVA School of Science and Technology, 2022 (cit. on pp. 18, 20).
- [72] I. Curioso et al. “Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-based Imputation”. In: *Journal of King Saud University - Computer and Information Sciences* 35.6 (2023-06-01), p. 101562. ISSN: 1319-1578. DOI: [10.1016/j.jksuci.2023.101562](https://doi.org/10.1016/j.jksuci.2023.101562) (cit. on pp. 70, 75, 90, 99, 163).
- [73] I. Curioso et al. “Pattern Recognition and Classification of Low-Intensity Emotions from Physiological Data”. In: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). L'Aquila, Italy: IEEE, 2023-06, pp. 299–304. ISBN: 9798350312249. DOI: [10.1109/CBMS58004.2023.00234](https://doi.org/10.1109/CBMS58004.2023.00234) (cit. on p. 166).
- [74] S. Das et al. “Precision Oncology in the Era of Genomics and Artificial Intelligence”. In: *Journal of Current Oncological Trends* 1.1 (2024-06), p. 22. ISSN: 4444-5555. DOI: [10.4103/JCOT.JCOT_3_23](https://doi.org/10.4103/JCOT.JCOT_3_23) (cit. on p. 33).
- [75] T. Davenport and R. Kalakota. “The potential for artificial intelligence in healthcare”. In: *Future Healthcare Journal* 6.2 (2019-06), pp. 94–98. ISSN: 2514-6645. DOI: [10.7861/futurehosp.6-2-94](https://doi.org/10.7861/futurehosp.6-2-94) (cit. on p. 33).
- [76] J. Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019-05-24. arXiv: [1810.04805\[cs\]](https://arxiv.org/abs/1810.04805) (cit. on pp. 16, 133, 135, 137, 146, 151, 225, 226).
- [77] M. Di Cesare et al. *World Heart Report 2023: Confronting the World's Number One Killer*. Geneva, Switzerland: World Heart Federation, 2023. URL: <https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf> (visited on 2024-09-29) (cit. on p. 111).
- [78] T. Dietterich. “Overfitting and undercomputing in machine learning”. In: *ACM Computing Surveys* 27.3 (1995-09), pp. 326–327. ISSN: 0360-0300, 1557-7341. DOI: [10.1145/212094.212114](https://doi.org/10.1145/212094.212114) (cit. on p. 12).
- [79] V. Dignum. “Responsible Artificial Intelligence: Designing AI for Human Values”. In: 1 (2017). URL: <https://www.itu.int/en/journal/001/Documents/itu2017-1.pdf> (visited on 2024-09-29) (cit. on p. 41).

- [80] A. Diker et al. "A novel application based on spectrogram and convolutional neural network for ECG classification". In: *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. IEEE, 2019, pp. 1–6. DOI: [10.1109/ubmyk48245.2019.8965506](https://doi.org/10.1109/ubmyk48245.2019.8965506) (cit. on pp. 113, 117).
- [81] T. Durand, N. Mehrasa, and G. Mori. "Learning a Deep ConvNet for Multi-Label Classification With Partial Labels". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019-06, pp. 647–657. ISBN: 978-1-72813-293-8. DOI: [10.1109/CVPR.2019.00074](https://doi.org/10.1109/CVPR.2019.00074) (cit. on pp. 89, 90).
- [82] B. Efron. "Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve". In: *Journal of the American Statistical Association* 83.402 (1988-06-01), pp. 414–425. ISSN: 0162-1459. DOI: [10.1080/01621459.1988.10478612](https://doi.org/10.1080/01621459.1988.10478612) (cit. on p. 131).
- [83] C. A. Efthymiou and D. J. O'Regan. "Postdischarge complications: what exactly happens when the patient goes home?" In: *Interactive CardioVascular and Thoracic Surgery* 12.2 (2011-02), pp. 130–134. ISSN: 1569-9285, 1569-9293. DOI: [10.1510/icvts.2010.249474](https://doi.org/10.1510/icvts.2010.249474) (cit. on pp. 48, 54).
- [84] B. Eini-Porat et al. "Tell me something interesting: Clinical utility of machine learning prediction models in the ICU". In: *Journal of Biomedical Informatics* 132 (2022-08), p. 104107. ISSN: 15320464. DOI: [10.1016/j.jbi.2022.104107](https://doi.org/10.1016/j.jbi.2022.104107) (cit. on p. 134).
- [85] M. A. El Mrabet, K. El Makkaoui, and A. Faize. "Supervised Machine Learning: A Survey". In: *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)*. 2021-12, pp. 1–10. DOI: [10.1109/CommNet52204.2021.9641998](https://doi.org/10.1109/CommNet52204.2021.9641998) (cit. on p. 11).
- [86] E. J. Emanuel. "What Makes Clinical Research Ethical?" In: *JAMA* 283.20 (2000-05-24), p. 2701. ISSN: 0098-7484. DOI: [10.1001/jama.283.20.2701](https://doi.org/10.1001/jama.283.20.2701) (cit. on p. 22).
- [87] T. Emmanuel et al. "A survey on missing data in machine learning". In: *Journal of Big Data* 8.1 (2021-10-27), p. 140. ISSN: 2196-1115. DOI: [10.1186/s40537-021-00516-9](https://doi.org/10.1186/s40537-021-00516-9) (cit. on p. 70).
- [88] C. K. Enders. *Applied missing data analysis, 2nd ed.* Applied missing data analysis, 2nd ed. Pages: ix, 546. New York, NY, US: The Guilford Press, 2022. ix, 546. ISBN: 978-1-4625-4986-3 (cit. on pp. 19, 71).
- [89] P. Esmailzadeh. "Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations". In: *Artificial Intelligence in Medicine* 151 (2024-05-01), p. 102861. ISSN: 0933-3657. DOI: [10.1016/j.artmed.2024.102861](https://doi.org/10.1016/j.artmed.2024.102861) (cit. on p. 41).

- [90] A. Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639 (2017-02). Publisher: Nature Publishing Group, pp. 115–118. ISSN: 1476-4687. DOI: [10.1038/nature21056](https://doi.org/10.1038/nature21056) (cit. on p. 30).
- [91] European Commission. Directorate General for Health and Food Safety. *European health data space*. LU: Publications Office, 2024. URL: <https://data.europa.eu/doi/10.2875/269514> (visited on 2024-09-02) (cit. on pp. 37, 161).
- [92] European Parliament and Council. *Artificial Intelligence Act*. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL_202401689 (visited on 2024-09-03) (cit. on pp. 28, 44).
- [93] European Parliament and Council of the European Union. *General Data Protection Regulation*. 2016. URL: <https://data.europa.eu/eli/reg/2016/679/oj> (visited on 2024-09-03) (cit. on pp. 38, 44).
- [94] A. Fallahpour et al. *EHRMamba: Towards Generalizable and Scalable Foundation Models for Electronic Health Records*. 2024-05-23. DOI: [10.48550/arXiv.2405.14567](https://doi.org/10.48550/arXiv.2405.14567). arXiv: [2405.14567\[cs\]](https://arxiv.org/abs/2405.14567) (cit. on p. 136).
- [95] M. Faltys et al. *HiRID, a high time-resolution ICU dataset (version 1.1.1)*. Version 1.1.1. 2021. DOI: [10.13026/NKWC-JS72](https://doi.org/10.13026/NKWC-JS72) (cit. on pp. 135, 141).
- [96] Y. Fan et al. “Development of machine learning models for mortality risk prediction after cardiac surgery”. In: *Cardiovascular Diagnosis and Therapy* 12.1 (2022-02), pp. 12–23. ISSN: 2223-3652. DOI: [10.21037/cdt-21-648](https://doi.org/10.21037/cdt-21-648) (cit. on p. 56).
- [97] F. A. C. D. Farias et al. “Remote Patient Monitoring: A Systematic Review”. In: *Telemedicine and e-Health* 26.5 (2020-05-01), pp. 576–583. ISSN: 1530-5627, 1556-3669. DOI: [10.1089/tmj.2019.0066](https://doi.org/10.1089/tmj.2019.0066) (cit. on p. 50).
- [98] H. Fei et al. “Topic-Enhanced Capsule Network for Multi-Label Emotion Classification”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1839–1848. ISSN: 2329-9290, 2329-9304. DOI: [10.1109/TASLP.2020.3001390](https://doi.org/10.1109/TASLP.2020.3001390) (cit. on p. 88).
- [99] M. P. B. Fernandes et al. “Machine Learning Models with Preoperative Risk Factors and Intraoperative Hypotension Parameters Predict Mortality After Cardiac Surgery”. In: *Journal of Cardiothoracic and Vascular Anesthesia* 35.3 (2021-03), pp. 857–865. ISSN: 10530770. DOI: [10.1053/j.jvca.2020.07.029](https://doi.org/10.1053/j.jvca.2020.07.029) (cit. on p. 56).
- [100] F. G. Fernandez et al. “The Society of Thoracic Surgeons National Database 2019 Annual Report”. In: *The Annals of Thoracic Surgery* 108.6 (2019-12), pp. 1625–1632. ISSN: 00034975. DOI: [10.1016/j.athoracsur.2019.09.034](https://doi.org/10.1016/j.athoracsur.2019.09.034) (cit. on p. 55).
- [101] P. L. Ferreira. “Development of the Portuguese version of MOS SF-36. Part I. Cultural and linguistic adaptation”. In: *Acta Medica Portuguesa* 13.1 (2000), pp. 55–66. ISSN: 0870-399X (cit. on p. 52).

- [102] G. M. Fitzmaurice and C. Ravichandran. “A Primer in Longitudinal Data Analysis”. In: *Circulation* 118.19 (2008-11-04), pp. 2005–2010. ISSN: 0009-7322, 1524-4539. DOI: [10.1161/CIRCULATIONAHA.107.714618](https://doi.org/10.1161/CIRCULATIONAHA.107.714618) (cit. on pp. 36, 130).
- [103] D. Folgado et al. “Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series”. In: *Information Fusion* 100 (2023-12-01), p. 101955. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2023.101955](https://doi.org/10.1016/j.inffus.2023.101955) (cit. on pp. 8, 40, 165).
- [104] D. E. Fry et al. “Inpatient and 90-Day Postdischarge Outcomes in Cardiac Surgery”. In: *American Journal of Account Care* 4.1 (2016), pp. 10–15. URL: [https://ajmc.s3.amazonaws.com/_media/_pdf/05_AJAC_Fry316%20\(final\).pdf](https://ajmc.s3.amazonaws.com/_media/_pdf/05_AJAC_Fry316%20(final).pdf) (visited on 2024-09-29) (cit. on p. 66).
- [105] Q. Fu et al. “Multi-Stations’ Weather Prediction Based on Hybrid Model Using 1D CNN and Bi-LSTM”. In: *2019 Chinese Control Conference (CCC)* (2019), pp. 3771–3775. DOI: [10.23919/ChiCC.2019.8866496](https://doi.org/10.23919/ChiCC.2019.8866496) (cit. on p. 113).
- [106] J. Gama et al. “A survey on concept drift adaptation”. In: *ACM Computing Surveys* 46.4 (2014-04). Publisher: Association for Computing Machinery (ACM), pp. 1–37. ISSN: 0360-0300, 1557-7341. DOI: [10.1145/2523813](https://doi.org/10.1145/2523813) (cit. on p. 26).
- [107] W. Ge et al. “An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units”. In: *Annual Symposium proceedings. AMIA Symposium 2018* (2018), pp. 460–469. ISSN: 1942-597X (cit. on pp. 31, 136).
- [108] Z. Ghahramani. “Unsupervised Learning”. In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. by O. Bousquet, U. von Luxburg, and G. Rätsch. Berlin, Heidelberg: Springer, 2004, pp. 72–112. ISBN: 978-3-540-28650-9. DOI: [10.1007/978-3-540-28650-9_5](https://doi.org/10.1007/978-3-540-28650-9_5) (cit. on p. 11).
- [109] M. Ghassemi et al. “A Review of Challenges and Opportunities in Machine Learning for Health”. In: *AMIA Summits on Translational Science Proceedings 2020* (2020-05-30), pp. 191–200. ISSN: 2153-4063 (cit. on pp. 3, 28, 31, 32, 35–37).
- [110] K. Ghosh et al. “The class imbalance problem in deep learning”. In: *Machine Learning* 113.7 (2024-07-01), pp. 4845–4901. ISSN: 1573-0565. DOI: [10.1007/s10994-022-06268-8](https://doi.org/10.1007/s10994-022-06268-8) (cit. on p. 70).
- [111] E. Gibaja and S. Ventura. “Multi-label learning: a review of the state of the art and ongoing research”. In: *WIREs Data Mining and Knowledge Discovery* 4.6 (2014-11), pp. 411–444. ISSN: 1942-4787, 1942-4795. DOI: [10.1002/widm.1139](https://doi.org/10.1002/widm.1139) (cit. on p. 88).

- [112] A. L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (2000-06-13). ISSN: 0009-7322, 1524-4539. DOI: [10.1161/01.CIR.101.23.e215](https://doi.org/10.1161/01.CIR.101.23.e215) (cit. on pp. 92, 134, 138).
- [113] Z. Golrizkhatami and A. Acan. "ECG classification using three-level fusion of different feature descriptors". In: *Expert Systems with Applications* 114 (2018-12), pp. 54–64. ISSN: 09574174. DOI: [10.1016/j.eswa.2018.07.030](https://doi.org/10.1016/j.eswa.2018.07.030). (Visited on 2023-11-22) (cit. on p. 114).
- [114] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. Adaptive computation and machine learning. Cambridge, Massachusetts London, England: The MIT Press, 2016. 775 pp. ISBN: 978-0-262-03561-3 (cit. on pp. 14, 15).
- [115] Y. Gorishniy, I. Rubachev, and A. Babenko. *On Embeddings for Numerical Features in Tabular Deep Learning*. 2023-10-26. arXiv: [2203.05556](https://arxiv.org/abs/2203.05556). URL: <http://arxiv.org/abs/2203.05556> (visited on 2024-11-07) (cit. on p. 157).
- [116] K. Greff et al. "LSTM: A Search Space Odyssey". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.10 (2017), pp. 2222–2232. DOI: [10.1109/TNNLS.2016.2582924](https://doi.org/10.1109/TNNLS.2016.2582924) (cit. on p. 15).
- [117] A. Gu and T. Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. 2023-12-01. DOI: [10.48550/arXiv.2312.00752](https://arxiv.org/abs/2312.00752). arXiv: [2312.00752\[cs\]](https://arxiv.org/abs/2312.00752). (Visited on 2024-01-08) (cit. on pp. 136, 156).
- [118] L. Guo, G. Sim, and B. Matuszewski. "Inter-patient ECG classification with convolutional and recurrent neural networks". In: *Biocybernetics and Biomedical Engineering* 39.3 (2019-07). Publisher: Elsevier Sp. z o.o., pp. 868–879. DOI: [10.1016/J.BBE.2019.06.001](https://doi.org/10.1016/J.BBE.2019.06.001) (cit. on p. 113).
- [119] I. Guyon and A. Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of machine learning research* (2003). URL: <https://dl.acm.org/doi/pdf/10.5555/944919.944968> (visited on 2024-09-29) (cit. on p. 23).
- [120] I. Habli, T. Lawton, and Z. Porter. "Artificial intelligence in health care: accountability and safety". In: *Bulletin of the World Health Organization* 98.4 (2020-04-01), pp. 251–256. ISSN: 0042-9686. DOI: [10.2471/BLT.19.237487](https://doi.org/10.2471/BLT.19.237487) (cit. on p. 32).
- [121] F. Hak, T. Guimarães, and M. Santos. "Towards effective clinical decision support systems: A systematic review". In: *PLoS ONE* 17.8 (2022-08-15), e0272846. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0272846](https://doi.org/10.1371/journal.pone.0272846) (cit. on p. 28).
- [122] M. Hammad, Y. Liu, and K. Wang. "Multimodal Biometric Authentication Systems Using Convolution Neural Network Based on Different Level Fusion of ECG and Fingerprint". In: *IEEE Access* 7 (2019), pp. 26527–26542. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2018.2886573](https://doi.org/10.1109/ACCESS.2018.2886573) (cit. on p. 114).

- [123] H. Han et al. "Multimodal multi-instance learning for long-term ECG classification". In: *Knowledge-Based Systems* 270 (2023-06), p. 110555. ISSN: 09507051. DOI: [10.1016/j.knosys.2023.110555](https://doi.org/10.1016/j.knosys.2023.110555) (cit. on p. 114).
- [124] H. M. Hapke and C. Nelson. *Building machine learning pipelines: automating model life cycles with TensorFlow*. First edition. Beijing Boston Farnham: O'Reilly, 2020. 337 pp. ISBN: 978-1-4920-5319-4 (cit. on pp. 21, 22).
- [125] M. A. Haque and S. Li. "Exploring ChatGPT and its impact on society". In: *AI and Ethics* (2024-02-21). ISSN: 2730-5961. DOI: [10.1007/s43681-024-00435-4](https://doi.org/10.1007/s43681-024-00435-4) (cit. on p. 3).
- [126] F. E. Harrell. "Evaluating the Yield of Medical Tests". In: *JAMA: The Journal of the American Medical Association* 247.18 (1982-05-14), p. 2543. ISSN: 0098-7484. DOI: [10.1001/jama.1982.03320430047030](https://doi.org/10.1001/jama.1982.03320430047030) (cit. on p. 64).
- [127] H. Harutyunyan et al. "Multitask learning and benchmarking with clinical time series data". In: *Scientific Data* 6.1 (2019-06-17). Number: 1 Publisher: Nature Publishing Group, p. 96. ISSN: 2052-4463. DOI: [10.1038/s41597-019-0103-9](https://doi.org/10.1038/s41597-019-0103-9) (cit. on pp. 149, 224).
- [128] S. J. Head et al. "The European Association for Cardio-Thoracic Surgery (EACTS) database: an introduction". In: *European Journal of Cardio-Thoracic Surgery* 44.3 (2013-09-01), e175–e180. ISSN: 1010-7940, 1873-734X. DOI: [10.1093/ejcts/ezt303](https://doi.org/10.1093/ejcts/ezt303) (cit. on pp. 51, 66).
- [129] S. A. Hicks et al. "On evaluation metrics for medical applications of artificial intelligence". In: *Scientific Reports* 12.1 (2022-04-08). Publisher: Nature Publishing Group, p. 5979. ISSN: 2045-2322. DOI: [10.1038/s41598-022-09954-8](https://doi.org/10.1038/s41598-022-09954-8) (cit. on p. 24).
- [130] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. European Commission, 2019. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (visited on 2024-09-03) (cit. on pp. 28, 38, 40, 42, 162).
- [131] S. Hirji et al. "Utility of 90-Day Mortality vs 30-Day Mortality as a Quality Metric for Transcatheter and Surgical Aortic Valve Replacement Outcomes". In: *JAMA Cardiology* 5.2 (2020-02-01), p. 156. ISSN: 2380-6583. DOI: [10.1001/jamacardio.2019.4657](https://doi.org/10.1001/jamacardio.2019.4657) (cit. on p. 66).
- [132] J. Hirsch et al. "ICD-10: History and Context". In: *American Journal of Neuroradiology* 37.4 (2016-04), pp. 596–599. ISSN: 0195-6108, 1936-959X. DOI: [10.3174/ajnr.A4696](https://doi.org/10.3174/ajnr.A4696) (cit. on p. 134).
- [133] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997-11-01), pp. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735) (cit. on pp. 112, 131, 154).

- [134] E. Hossain et al. "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review". In: *Computers in Biology and Medicine* 155 (2023-03-01), p. 106649. ISSN: 0010-4825. DOI: [10.1016/j.combiomed.2023.106649](https://doi.org/10.1016/j.combiomed.2023.106649) (cit. on pp. 11, 30).
- [135] J. Huang et al. "ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network". In: *IEEE Access* 7 (2019), pp. 92871–92880. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2928017](https://doi.org/10.1109/ACCESS.2019.2928017) (cit. on p. 113).
- [136] J. Huang et al. "Improving multi-label classification with missing labels by learning label-specific features". In: *Information Sciences* 492 (2019-08), pp. 124–146. ISSN: 00200255. DOI: [10.1016/j.ins.2019.04.021](https://doi.org/10.1016/j.ins.2019.04.021) (cit. on pp. 89, 90).
- [137] J. Huang et al. "Multi-label learning with missing and completely unobserved labels". In: *Data Mining and Knowledge Discovery* 35.3 (2021-05), pp. 1061–1086. ISSN: 1384-5810, 1573-756X. DOI: [10.1007/s10618-021-00743-x](https://doi.org/10.1007/s10618-021-00743-x) (cit. on pp. 89, 90).
- [138] S.-C. Huang et al. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines". In: *NPJ digital medicine* 3.1 (2020). Publisher: Nature Publishing Group, pp. 1–9. DOI: [10.1038/s41746-020-00341-z](https://doi.org/10.1038/s41746-020-00341-z) (cit. on pp. 109, 110, 115, 118).
- [139] K. Hur et al. *UniHPF : Universal Healthcare Predictive Framework with Zero Domain Knowledge*. 2022-11-15. arXiv: [2211.08082\[cs\]](https://arxiv.org/abs/2211.08082) (cit. on p. 130).
- [140] S. L. Hyland et al. "Early prediction of circulatory failure in the intensive care unit using machine learning". In: *Nature Medicine* 26.3 (2020-03), pp. 364–373. DOI: [10.1038/s41591-020-0789-4](https://doi.org/10.1038/s41591-020-0789-4) (cit. on p. 141).
- [141] K. M. Ibrahim et al. "Confidence-based Weighted Loss for Multi-label Classification with Missing Labels". In: *Proceedings of the 2020 International Conference on Multimedia Retrieval*. ICMR '20: International Conference on Multimedia Retrieval. Dublin Ireland: ACM, 2020-06-08, pp. 291–295. ISBN: 978-1-4503-7087-5. DOI: [10.1145/3372278.3390728](https://doi.org/10.1145/3372278.3390728) (cit. on pp. 89, 90).
- [142] IMDRF SaMD Working Group. *Software as a Medical Device (SaMD): Key Definitions*. 2013. URL: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf> (visited on 2024-09-03) (cit. on p. 46).
- [143] A. Iranfar, A. Arza, and D. Atienza. *ReLearn: A Robust Machine Learning Framework in Presence of Missing Data for Multimodal Stress Detection from Physiological Signals*. 2021-07-29. arXiv: [2104.14278\[cs\]](https://arxiv.org/abs/2104.14278) (cit. on p. 70).
- [144] S. Ivic. "The Concept of European Values". In: *Cultura* 16.1 (2019-01-01), pp. 103–117. ISSN: 1584-1057. DOI: [10.3726/CUL012019.0007](https://doi.org/10.3726/CUL012019.0007) (cit. on p. 42).

- [145] R. Jain, S. Duval, and S. Adabag. “How Accurate Is the Eyeball Test?: A Comparison of Physician’s Subjective Assessment Versus Statistical Methods in Estimating Mortality Risk After Cardiac Surgery”. In: *Circulation: Cardiovascular Quality and Outcomes* 7.1 (2014-01), pp. 151–156. ISSN: 1941-7713, 1941-7705. DOI: [10.1161/CIRCOUTCOMES.113.000329](https://doi.org/10.1161/CIRCOUTCOMES.113.000329) (cit. on p. 67).
- [146] V. Jain, N. Modhe, and P. Rai. “Scalable Generative Models for Multi-label Learning with Missing Labels”. In: *International Conference on Machine Learning*. 2017. URL: <https://dl.acm.org/doi/pdf/10.5555/3305381.3305550> (visited on 2024-09-29) (cit. on pp. 89, 90).
- [147] F. Jiang et al. “Artificial intelligence in healthcare: past, present and future”. In: *Stroke and Vascular Neurology* 2.4 (2017-12), pp. 230–243. ISSN: 2059-8688, 2059-8696. DOI: [10.1136/svn-2017-000101](https://doi.org/10.1136/svn-2017-000101) (cit. on p. 29).
- [148] A. E. W. Johnson et al. “The MIMIC Code Repository: enabling reproducibility in critical care research”. In: *Journal of the American Medical Informatics Association* 25.1 (2018-01-01), pp. 32–39. ISSN: 1527-974X. DOI: [10.1093/jamia/ocx084](https://doi.org/10.1093/jamia/ocx084) (cit. on pp. 139, 142).
- [149] A. E. W. Johnson et al. “Machine Learning and Decision Support in Critical Care”. In: *Proceedings of the IEEE* 104.2 (2016-02). Conference Name: Proceedings of the IEEE, pp. 444–466. ISSN: 1558-2256. DOI: [10.1109/JPROC.2015.2501978](https://doi.org/10.1109/JPROC.2015.2501978) (cit. on pp. 3, 31, 36–38, 134).
- [150] A. E. W. Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific Data* 3.1 (2016-05-24). Publisher: Nature Publishing Group, p. 160035. ISSN: 2052-4463. DOI: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35) (cit. on pp. 135, 140).
- [151] A. E. W. Johnson et al. “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific Data* 10.1 (2023-01-03), p. 1. ISSN: 2052-4463. DOI: [10.1038/s41597-022-01899-x](https://doi.org/10.1038/s41597-022-01899-x) (cit. on pp. 134, 138).
- [152] K. W. Johnson et al. “Artificial Intelligence in Cardiology”. In: *Journal of the American College of Cardiology* 71.23 (2018-06), pp. 2668–2679. ISSN: 07351097. DOI: [10.1016/j.jacc.2018.03.521](https://doi.org/10.1016/j.jacc.2018.03.521) (cit. on p. 56).
- [153] D. Jonkers. “Prevalence of 90-days postoperative wound infections after cardiac surgery”. In: *European Journal of Cardio-Thoracic Surgery* 23.1 (2003-01), pp. 97–102. ISSN: 10107940. DOI: [10.1016/S1010-7940\(02\)00662-0](https://doi.org/10.1016/S1010-7940(02)00662-0) (cit. on p. 66).
- [154] L. P. Kaelbling, M. L. Littman, and A. W. Moore. “Reinforcement Learning: A Survey”. In: *Journal of Artificial Intelligence Research* 4 (1996-05-01), pp. 237–285. ISSN: 1076-9757. DOI: [10.1613/jair.301](https://doi.org/10.1613/jair.301) (cit. on p. 11).

- [155] M. N. Kamel Boulos and P. Zhang. “Digital Twins: From Personalised Medicine to Precision Public Health”. In: *Journal of Personalized Medicine* 11.8 (2021-08). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 745. ISSN: 2075-4426. DOI: [10.3390/jpm11080745](https://doi.org/10.3390/jpm11080745) (cit. on p. 33).
- [156] A. Katrompas, T. Ntakouris, and V. Metsis. “Recurrence and Self-attention vs the Transformer for Time-Series Classification: A Comparative Study”. In: *Artificial Intelligence in Medicine*. Cham: Springer International Publishing, 2022, pp. 99–109. ISBN: 978-3-031-09342-5. DOI: [10.1007/978-3-031-09342-5_10](https://doi.org/10.1007/978-3-031-09342-5_10) (cit. on p. 132).
- [157] E. Katsoulakis et al. “Digital twins for health: a scoping review”. In: *npj Digital Medicine* 7.1 (2024-03-22). Publisher: Nature Publishing Group, pp. 1–11. ISSN: 2398-6352. DOI: [10.1038/s41746-024-01073-0](https://doi.org/10.1038/s41746-024-01073-0) (cit. on p. 33).
- [158] KDIGO. “Section 2: AKI Definition”. In: *Kidney International Supplements*. KDIGO Clinical Practice Guideline for Acute Kidney Injury 2.1 (2012-03-01), pp. 19–36. ISSN: 2157-1716. DOI: [10.1038/kisup.2011.32](https://doi.org/10.1038/kisup.2011.32) (cit. on p. 149).
- [159] G. Ke et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: <https://dl.acm.org/doi/pdf/10.5555/3294996.3295074> (visited on 2024-08-07) (cit. on pp. 13, 154).
- [160] M. Kelly, R. Longjohn, and K. Nottingham. *The UCI Machine Learning Repository*. URL: <https://archive.ics.uci.edu> (cit. on p. 74).
- [161] A. S. Kesselheim et al. “Clinical Decision Support Systems Could Be Modified To Reduce ‘Alert Fatigue’ While Still Minimizing The Risk Of Litigation”. In: *Health Affairs* 30.12 (2011-12), pp. 2310–2317. ISSN: 0278-2715, 1544-5208. DOI: [10.1377/hlthaff.2010.1111](https://doi.org/10.1377/hlthaff.2010.1111) (cit. on p. 32).
- [162] S. Kessler et al. “Predicting readmission to the cardiovascular intensive care unit using recurrent neural networks”. In: *DIGITAL HEALTH* 9 (2023-01). ISSN: 2055-2076, 2055-2076. DOI: [10.1177/20552076221149529](https://doi.org/10.1177/20552076221149529) (cit. on p. 136).
- [163] B. Khan et al. “Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector”. In: *Biomedical Materials & Devices (New York, N.y.)* (2023-02-08), pp. 1–8. ISSN: 2731-4812. DOI: [10.1007/s44174-023-00063-2](https://doi.org/10.1007/s44174-023-00063-2). (Visited on 2024-08-30) (cit. on p. 32).
- [164] H. Khan, X. Wang, and H. Liu. “Handling missing data through deep convolutional neural network”. In: *Information Sciences* 595 (2022-05-01), pp. 278–293. ISSN: 0020-0255. DOI: [10.1016/j.ins.2022.02.051](https://doi.org/10.1016/j.ins.2022.02.051) (cit. on p. 72).
- [165] H. Khoury et al. “Readmission After Surgical Aortic Valve Replacement in the United States”. In: *The Annals of Thoracic Surgery* 110.3 (2020-09), pp. 849–855. ISSN: 1552-6259. DOI: [10.1016/j.athoracsur.2019.11.058](https://doi.org/10.1016/j.athoracsur.2019.11.058) (cit. on p. 48).

- [166] M. Kiener. “Artificial intelligence in medicine and the disclosure of risks”. In: *AI & SOCIETY* 36.3 (2021-09-01), pp. 705–713. ISSN: 1435-5655. DOI: [10.1007/s00146-020-01085-w](https://doi.org/10.1007/s00146-020-01085-w) (cit. on p. 32).
- [167] A. Kilic et al. “Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery”. In: *The Annals of Thoracic Surgery* 109.6 (2020-06), pp. 1811–1819. ISSN: 00034975. DOI: [10.1016/j.athoracsur.2019.09.049](https://doi.org/10.1016/j.athoracsur.2019.09.049) (cit. on p. 57).
- [168] G. King and L. Zeng. “Logistic Regression in Rare Events Data”. In: *Journal of Statistical Software* 8.2 (2003). ISSN: 1548-7660. DOI: [10.18637/jss.v008.i02](https://doi.org/10.18637/jss.v008.i02) (cit. on p. 60).
- [169] L. Koessler and J. Schuett. *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries*. 2023-07-17. arXiv: [2307.08823](https://arxiv.org/abs/2307.08823)[cs] (cit. on p. 45).
- [170] M. Komorowski et al. “The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care”. In: *Nature Medicine* 24.11 (2018-11), pp. 1716–1720. ISSN: 1078-8956, 1546-170X. DOI: [10.1038/s41591-018-0213-5](https://doi.org/10.1038/s41591-018-0213-5) (cit. on p. 31).
- [171] G. Kong, K. Lin, and Y. Hu. “Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU”. In: *BMC Medical Informatics and Decision Making* 20.1 (2020-10-02), p. 251. ISSN: 1472-6947. DOI: [10.1186/s12911-020-01271-2](https://doi.org/10.1186/s12911-020-01271-2) (cit. on p. 135).
- [172] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. New York, NY: Springer New York, 2013. DOI: [10.1007/978-1-4614-6849-3](https://doi.org/10.1007/978-1-4614-6849-3) (cit. on p. 22).
- [173] S. Kumar et al. “Potential Impact of Data-Centric AI on Society”. In: *IEEE Technology and Society Magazine* 42.3 (2023-09). Conference Name: IEEE Technology and Society Magazine, pp. 98–107. ISSN: 1937-416X. DOI: [10.1109/MTS.2023.3306532](https://doi.org/10.1109/MTS.2023.3306532). (Visited on 2024-09-03) (cit. on p. 40).
- [174] A. Labach et al. *DuETT: Dual Event Time Transformer for Electronic Health Records*. 2023-08-15. arXiv: [2304.13017](https://arxiv.org/abs/2304.13017)[cs] (cit. on pp. 134, 136, 148, 149, 156).
- [175] J. Lai et al. “Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset”. In: *Nature Communications* 14.1 (2023). Publisher: Nature Publishing Group UK London, p. 3741. DOI: [10.1038/s41467-023-39472-8](https://doi.org/10.1038/s41467-023-39472-8) (cit. on p. 113).
- [176] J. Laitala et al. “Robust ECG R-peak Detection Using LSTM”. In: *SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing*. Vol. 20. 2020, pp. 1104–1111. ISBN: 978-1-4503-6866-7. DOI: [10.1145/3341105.3373945](https://doi.org/10.1145/3341105.3373945) (cit. on p. 113).

- [177] S. I. Lambert et al. “An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals”. In: *npj Digital Medicine* 6.1 (2023-06-10). Publisher: Nature Publishing Group, pp. 1–14. ISSN: 2398-6352. DOI: [10.1038/s41746-023-00852-5](https://doi.org/10.1038/s41746-023-00852-5) (cit. on p. 32).
- [178] C. Lea et al. *Temporal Convolutional Networks: A Unified Approach to Action Segmentation*. 2016-08-29. DOI: [10.48550/arXiv.1608.08242](https://doi.org/10.48550/arXiv.1608.08242). arXiv: [1608.08242\[cs\]](https://arxiv.org/abs/1608.08242) (cit. on p. 154).
- [179] Y. LeCun, Y. Bengio, and G. Hinton. “Deep learning”. In: *Nature* 521.7553 (2015-05-28), pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539) (cit. on pp. 14, 15).
- [180] R. S. Ledley and L. B. Lusted. “Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason”. In: *Science (New York, N.Y.)* 130.3366 (1959-07-03), pp. 9–21. ISSN: 0036-8075. DOI: [10.1126/science.130.3366.9](https://doi.org/10.1126/science.130.3366.9) (cit. on p. 29).
- [181] J. Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020-02-15), pp. 1234–1240. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682). (Visited on 2024-01-08) (cit. on pp. 137, 225).
- [182] M. Lentzen et al. “ExMed-BERT: A Transformer-Based Model Trained on Large Scale Claims Data for Prediction of Severe COVID-19 Disease Progression”. In: *IEEE Journal of Biomedical and Health Informatics* 27.9 (2023-09), pp. 4548–4558. ISSN: 2168-2194, 2168-2208. DOI: [10.1109/JBHI.2023.3288768](https://doi.org/10.1109/JBHI.2023.3288768) (cit. on pp. 130, 136).
- [183] C. Li et al. *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day*. 2023-06-01. arXiv: [2306.00890\[cs\]](https://arxiv.org/abs/2306.00890) (cit. on p. 128).
- [184] F. Li et al. “Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database”. In: *BMJ Open* 11.7 (2021-07-01). Publisher: British Medical Journal Publishing Group Section: Cardiovascular medicine, e044779. ISSN: 2044-6055, 2044-6055. DOI: [10.1136/bmjopen-2020-044779](https://doi.org/10.1136/bmjopen-2020-044779) (cit. on p. 135).
- [185] H. Z. Li and P. Boulanger. “A Survey of Heart Anomaly Detection Using Ambulatory Electrocardiogram (ECG)”. In: *Sensors (Basel, Switzerland)* 20.5 (2020-03). Publisher: Multidisciplinary Digital Publishing Institute (MDPI). ISSN: 14248220. DOI: [10.3390/S20051461](https://doi.org/10.3390/S20051461) (cit. on p. 116).
- [186] Y. Li et al. “BEHRT: Transformer for Electronic Health Records”. In: *Scientific Reports* 10.1 (2020-04-28), p. 7155. ISSN: 2045-2322. DOI: [10.1038/s41598-020-62922-y](https://doi.org/10.1038/s41598-020-62922-y) (cit. on pp. 134, 136).

-
- [187] Y. Li et al. *Hi-BEHRT: Hierarchical Transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records*. 2021-06-21. arXiv: [2106.11360\[cs\]](#) (cit. on p. 136).
 - [188] P. P. Liang, A. Zadeh, and L.-P. Morency. “Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions”. In: *ACM Comput. Surv.* 56.10 (2024-06-22), 264:1–264:42. ISSN: 0360-0300. DOI: [10.1145/3656580](#) (cit. on pp. 107–109).
 - [189] M. Lindstrom et al. “Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021”. In: *Journal of the American College of Cardiology* 80.25 (2022-12), pp. 2372–2425. ISSN: 07351097. DOI: [10.1016/j.jacc.2022.11.001](#) (cit. on p. 106).
 - [190] G. Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42 (2017-12), pp. 60–88. ISSN: 13618415. DOI: [10.1016/j.media.2017.07.005](#) (cit. on pp. 29, 30).
 - [191] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Google-Books-ID: BemMDwAAQBAJ. John Wiley & Sons, 2019-04-23. 462 pp. ISBN: 978-0-470-52679-8 (cit. on pp. 18, 37, 91, 93).
 - [192] C.-F. Liu et al. “The Critical Factors Affecting the Deployment and Scaling of Healthcare AI: Viewpoint from an Experienced Medical Center”. In: *Healthcare* 9.6 (2021-06). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 685. ISSN: 2227-9032. DOI: [10.3390/healthcare9060685](#) (cit. on p. 41).
 - [193] F. Liu et al. “An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection”. In: *Journal of Medical Imaging and Health Informatics* 8.7 (2018-09-01), pp. 1368–1373. ISSN: 2156-7018. DOI: [10.1166/jmih.2018.2442](#) (cit. on p. 92).
 - [194] L. Liu et al. *Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction*. 2018-11-17. arXiv: [1803.04837\[cs, stat\]](#) (cit. on p. 136).
 - [195] M. Liu and Y. Kim. “Classification of Heart Diseases Based on ECG Signals Using Long Short-Term Memory”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc., 2018-10, pp. 2707–2710. DOI: [10.1109/EMBC.2018.8512761](#) (cit. on p. 113).
 - [196] X. Liu, X. Lai, and L. Zhang. “A Hierarchical Missing Value Imputation Method by Correlation-Based K-Nearest Neighbors”. In: *Intelligent Systems and Applications*. Cham: Springer International Publishing, 2019, pp. 486–496. ISBN: 978-3-030-29516-5. DOI: [10.1007/978-3-030-29516-5_38](#) (cit. on pp. 72, 73, 76, 80, 91).

- [197] X. Liu et al. "Deep learning in ECG diagnosis: A review". In: *Knowledge-Based Systems* 227 (2021). Publisher: Elsevier, p. 107187. DOI: [10.1016/j.knsys.2021.107187](https://doi.org/10.1016/j.knsys.2021.107187) (cit. on pp. 111, 113).
- [198] A. Londral et al. "Developing and validating high-value patient digital follow-up services: a pilot study in cardiac surgery". In: *BMC Health Services Research* 22.1 (2022-05-21), p. 680. ISSN: 1472-6963. DOI: [10.1186/s12913-022-08073-4](https://doi.org/10.1186/s12913-022-08073-4) (cit. on p. 50).
- [199] I. Lopes et al. "Smartphone-Based Remote Monitoring Solution for Heart Failure Patients". In: *Studies in Health Technology and Informatics* 261 (2019), pp. 109–114. ISSN: 1879-8365 (cit. on p. 50).
- [200] J. M. Lourenço. *The NOVAthesis L^AT_EX Template User's Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/main/template.pdf> (cit. on p. i).
- [201] J. Lu et al. "Learning under Concept Drift: A Review". In: *IEEE Transactions on Knowledge and Data Engineering* (2018), pp. 1–1. ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: [10.1109/TKDE.2018.2876857](https://doi.org/10.1109/TKDE.2018.2876857). arXiv: [2004.05785\[cs, stat\]](https://arxiv.org/abs/2004.05785) (cit. on pp. 26, 32).
- [202] S. Lundberg and S.-I. Lee. *A Unified Approach to Interpreting Model Predictions*. 2017-11-24. DOI: [10.48550/arXiv.1705.07874](https://doi.org/10.48550/arXiv.1705.07874). arXiv: [1705.07874\[cs, stat\]](https://arxiv.org/abs/1705.07874) (cit. on p. 66).
- [203] J. Ma et al. "Segment Anything in Medical Images". In: *Nature Communications* 15.1 (2024-01-22), p. 654. ISSN: 2041-1723. DOI: [10.1038/s41467-024-44824-z](https://doi.org/10.1038/s41467-024-44824-z). arXiv: [2304.12306\[cs, eess\]](https://arxiv.org/abs/2304.12306) (cit. on pp. 10, 30).
- [204] M. Ma et al. *SMIL: Multimodal Learning with Severely Missing Modality*. 2021-03-09. DOI: [10.48550/arXiv.2103.05677](https://doi.org/10.48550/arXiv.2103.05677). arXiv: [2103.05677\[cs\]](https://arxiv.org/abs/2103.05677) (cit. on p. 128).
- [205] A. Majeed and S. O. Hwang. "Technical Analysis of Data-Centric and Model-Centric Artificial Intelligence". In: *IT Professional* 25.6 (2023-11). Conference Name: IT Professional, pp. 62–70. ISSN: 1941-045X. DOI: [10.1109/MITP.2023.3322410](https://doi.org/10.1109/MITP.2023.3322410) (cit. on p. 40).
- [206] G. Manias et al. "iHELP: Personalised Health Monitoring and Decision Support Based on Artificial Intelligence and Holistic Health Records". In: *2021 IEEE Symposium on Computers and Communications (ISCC)*. 2021 IEEE Symposium on Computers and Communications (ISCC). Athens, Greece: IEEE, 2021-09-05, pp. 1–8. ISBN: 978-1-66542-744-9. DOI: [10.1109/ISCC53001.2021.9631475](https://doi.org/10.1109/ISCC53001.2021.9631475) (cit. on p. 33).
- [207] L. T. Martin et al. "The Issues of Interoperability and Data Connectedness for Public Health". In: *Big Data* 10 (Suppl 1 2022-09-01), S19–S24. ISSN: 2167-6461. DOI: [10.1089/big.2022.0207](https://doi.org/10.1089/big.2022.0207) (cit. on p. 37).

- [208] G. Mastroiacovo et al. “Is EuroSCORE II still a reliable predictor for cardiac surgery mortality in 2022? A retrospective study study”. In: *European Journal of Cardio-Thoracic Surgery* 64.3 (2022-01-01), ezad294. ISSN: 1873-734X. DOI: [10.1093/ejcts/ezad294](https://doi.org/10.1093/ejcts/ezad294) (cit. on p. 49).
- [209] P. Mathur et al. “Artificial Intelligence, Machine Learning, and Cardiovascular Disease”. In: *Clinical Medicine Insights: Cardiology* 14 (2020-01-01). Publisher: SAGE Publications Ltd STM, p. 1179546820927404. ISSN: 1179-5468. DOI: [10.1177/1179546820927404](https://doi.org/10.1177/1179546820927404) (cit. on p. 33).
- [210] P. Matias et al. “Time Series Segmentation Using Neural Networks with Cross-Domain Transfer Learning”. In: *Electronics* 10.15 (2021-07-28), p. 1805. ISSN: 2079-9292. DOI: [10.3390/electronics10151805](https://doi.org/10.3390/electronics10151805) (cit. on p. 113).
- [211] D. McGraw et al. “Privacy As An Enabler, Not An Impediment: Building Trust Into Health Information Exchange”. In: *Health Affairs* 28.2 (2009-03), pp. 416–427. ISSN: 0278-2715, 1544-5208. DOI: [10.1377/hlthaff.28.2.416](https://doi.org/10.1377/hlthaff.28.2.416) (cit. on p. 38).
- [212] S. McLean et al. “The risks associated with Artificial General Intelligence: A systematic review”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 35.5 (2023-07-04), pp. 649–663. ISSN: 0952-813X, 1362-3079. DOI: [10.1080/0952813X.2021.1964003](https://doi.org/10.1080/0952813X.2021.1964003) (cit. on p. 45).
- [213] G. Medic et al. “Evidence-based Clinical Decision Support Systems for the prediction and detection of three disease states in critical care: A systematic literature review”. In: *F1000Research* 8 (2019-11-27), p. 1728. ISSN: 2046-1402. DOI: [10.12688/f1000research.20498.2](https://doi.org/10.12688/f1000research.20498.2) (cit. on p. 29).
- [214] Medical Device Coordination Group Document. *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*. 2019. URL: https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2019_11_guidance_qualification_classification_software_en_0.pdf (visited on 2024-09-03) (cit. on p. 46).
- [215] Y. Meng et al. “BRLTM: Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression”. In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021-08), pp. 3121–3129. ISSN: 2168-2194, 2168-2208. DOI: [10.1109/JBHI.2021.3063721](https://doi.org/10.1109/JBHI.2021.3063721) (cit. on p. 136).
- [216] S. M. Meystre et al. “Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress”. In: *Yearbook of Medical Informatics* 26.1 (2017), pp. 38–52. ISSN: 0943-4747, 2364-0502. DOI: [10.15265/IY-2017-007](https://doi.org/10.15265/IY-2017-007) (cit. on pp. 3, 35, 134).
- [217] B. Middleton, D. F. Sittig, and A. Wright. “Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision”. In: *Yearbook of Medical Informatics Suppl 1* (Suppl 1 2016-08-02), S103–116. ISSN: 2364-0502. DOI: [10.15265/IYS-2016-s034](https://doi.org/10.15265/IYS-2016-s034) (cit. on p. 30).

- [218] A. Minchol  et al. "Machine learning in the electrocardiogram". In: *Journal of Electrocardiology* 57 (2019-11), S61–S64. ISSN: 00220736. DOI: [10.1016/j.jelectrocard.2019.08.008](https://doi.org/10.1016/j.jelectrocard.2019.08.008) (cit. on p. 111).
- [219] P. Mishra et al. *FCMI: Feature Correlation based Missing Data Imputation*. 2021-06-26. arXiv: [2107.00100\[cs\]](https://arxiv.org/abs/2107.00100) (cit. on pp. 72, 73).
- [220] S. Moazemi et al. "Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review". In: *Frontiers in Medicine* 10 (2023-03-31), p. 1109411. ISSN: 2296-858X. DOI: [10.3389/fmed.2023.1109411](https://doi.org/10.3389/fmed.2023.1109411) (cit. on pp. 134, 135).
- [221] Mohebbanaaz, Y. P. Sai, and L. R. Kumari. "A Review on Arrhythmia Classification Using ECG Signals". In: *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. Bhopal, India: IEEE, 2020-02, pp. 1–6. ISBN: 978-1-72814-862-5. DOI: [10.1109/SCEECS48394.2020.9](https://doi.org/10.1109/SCEECS48394.2020.9) (cit. on p. 92).
- [222] V. Mond jar-Guerra et al. "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers". In: *Biomedical Signal Processing and Control* 47 (2019-01), pp. 41–48. ISSN: 17468094. DOI: [10.1016/j.bspc.2018.08.007](https://doi.org/10.1016/j.bspc.2018.08.007) (cit. on p. 92).
- [223] B. J. Mortazavi et al. "Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures". In: *IEEE Journal of Biomedical and Health Informatics* 21.6 (2017-11), pp. 1719–1729. ISSN: 2168-2194, 2168-2208. DOI: [10.1109/JBHI.2017.2675340](https://doi.org/10.1109/JBHI.2017.2675340) (cit. on pp. 31, 56, 66).
- [224] L. Mou et al. "T-LSTM: A Long Short-Term Memory Neural Network Enhanced by Temporal Information for Traffic Flow Prediction". In: *IEEE Access* 7 (2019), pp. 98053–98060. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2929692](https://doi.org/10.1109/ACCESS.2019.2929692) (cit. on pp. 132, 136).
- [225] K. P. Murphy. *Probabilistic machine learning: an introduction*. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2022. 826 pp. ISBN: 978-0-262-04682-4 (cit. on p. 13).
- [226] S. Murthy et al. "A Comparative Study of Data Anonymization Techniques". In: *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2019-05, pp. 306–309. ISBN: 978-1-72810-006-7. DOI: [10.1109/BigDataSecurity-HPSC-IDS.2019.00063](https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00063) (cit. on p. 22).
- [227] H. Narotamo et al. "Deep learning for ECG classification: A comparative study of 1D and 2D representations and multimodal fusion approaches". In: *Biomedical Signal Processing and Control* 93 (2024-07), p. 106141. ISSN: 17468094. DOI: [10.1016/j.bspc.2024.106141](https://doi.org/10.1016/j.bspc.2024.106141) (cit. on pp. 111, 165).

- [228] S. A. M. Nashef et al. "EuroSCORE II". In: *European Journal of Cardio-Thoracic Surgery* 41.4 (2012-04-01), pp. 734–745. ISSN: 1010-7940, 1873-734X. DOI: [10.1093/ejcts/ezs043](https://doi.org/10.1093/ejcts/ezs043) (cit. on pp. 48, 54, 55, 66).
- [229] S. Nashef et al. "European system for cardiac operative risk evaluation (EuroSCORE)". In: *European Journal of Cardio-Thoracic Surgery* 16.1 (1999-07), pp. 9–13. ISSN: 1873-734X, 1010-7940. DOI: [10.1016/S1010-7940\(99\)00134-7](https://doi.org/10.1016/S1010-7940(99)00134-7) (cit. on pp. 30, 55).
- [230] Z. Niu, G. Zhong, and H. Yu. "A review on the attention mechanism of deep learning". In: *Neurocomputing* 452 (2021-09-10), pp. 48–62. ISSN: 0925-2312. DOI: [10.1016/j.neucom.2021.03.091](https://doi.org/10.1016/j.neucom.2021.03.091) (cit. on p. 16).
- [231] C. Nwankpa et al. *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*. 2018-11-08. arXiv: [1811.03378\[cs\]](https://arxiv.org/abs/1811.03378) (cit. on p. 16).
- [232] S. M. O'Brien et al. "The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2—Statistical Methods and Results". In: *The Annals of Thoracic Surgery* 105.5 (2018-05), pp. 1419–1428. ISSN: 00034975. DOI: [10.1016/j.athoracsur.2018.03.003](https://doi.org/10.1016/j.athoracsur.2018.03.003) (cit. on pp. 49, 54).
- [233] Z. Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019-10-25). Publisher: American Association for the Advancement of Science, pp. 447–453. DOI: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342) (cit. on p. 38).
- [234] OECD. *Recommendation of the Council on Artificial Intelligence*. Version OECD/LEGAL/0449. 2024. URL: <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449> (visited on 2024-09-03) (cit. on pp. 43, 44).
- [235] OpenAI et al. *GPT-4 Technical Report*. 2024-03-04. arXiv: [2303.08774\[cs\]](https://arxiv.org/abs/2303.08774) (cit. on pp. 3, 16, 45, 109, 130, 133).
- [236] K. Pang et al. "Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database". In: *Diagnostics* 12.5 (2022-05). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 1068. ISSN: 2075-4418. DOI: [10.3390/diagnostics12051068](https://doi.org/10.3390/diagnostics12051068) (cit. on p. 135).
- [237] R. B. Parikh, S. Teeple, and A. S. Navathe. "Addressing Bias in Artificial Intelligence in Health Care". In: *JAMA* 322.24 (2019-12-24), p. 2377. ISSN: 0098-7484. DOI: [10.1001/jama.2019.18058](https://doi.org/10.1001/jama.2019.18058) (cit. on p. 38).
- [238] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. 2018-06-05. DOI: [10.48550/arXiv.1201.0490](https://doi.org/10.48550/arXiv.1201.0490). arXiv: [1201.0490\[cs\]](https://arxiv.org/abs/1201.0490) (cit. on p. 60).
- [239] R. Pekmezaris et al. "Home Telemonitoring In Heart Failure: A Systematic Review And Meta-Analysis". In: *Health Affairs* 37.12 (2018-12), pp. 1983–1989. ISSN: 0278-2715, 1544-5208. DOI: [10.1377/hlthaff.2018.05087](https://doi.org/10.1377/hlthaff.2018.05087) (cit. on p. 50).

- [240] M. Pinto et al. “The Matrix Reloaded: Towards Counterfactual Group Fairness in Machine Learning”. In: *Journal of Data-centric Machine Learning Research* (2024) (cit. on p. 33).
- [241] A. P. Pittams et al. “Scoring Systems for Risk Stratification in Patients Undergoing Cardiac Surgery”. In: *Journal of Cardiothoracic and Vascular Anesthesia* 36.4 (2022-04), pp. 1148–1156. ISSN: 10530770. DOI: [10.1053/j.jvca.2021.03.005](https://doi.org/10.1053/j.jvca.2021.03.005) (cit. on p. 54).
- [242] M. Pleasant. *The Most Important Truth. The Power of Your Own Voice. Challenging the Denial of the Climate Crisis and Our Hacked Demoratic Process*. In collab. with A. Gore. 2013-09-02. URL: https://issuu.com/originmagazine/docs/origin_14_art_conscious_02 (visited on 2024-09-06) (cit. on p. 167).
- [243] T. J. Pollard et al. “The eICU Collaborative Research Database, a freely available multi-center database for critical care research”. In: *Scientific Data* 5.1 (2018-09-11), p. 180178. ISSN: 2052-4463. DOI: [10.1038/sdata.2018.178](https://doi.org/10.1038/sdata.2018.178) (cit. on pp. 135, 140).
- [244] R. Poulain and R. Beheshti. “GT-BEHRT: Graph Transformers on EHRs: Better Representation Improves Downstream Performance”. In: The Twelfth International Conference on Learning Representations. 2023-10-13. URL: <https://openreview.net/forum?id=pe0Vdv7rsL> (cit. on p. 136).
- [245] D. M. W. Powers. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 2020-10-10. arXiv: [2010.16061\[cs,stat\]](https://arxiv.org/abs/2010.16061) (cit. on p. 24).
- [246] P. K. S. Prakash et al. “RareBERT: Transformer Architecture for Rare Disease Patient Identification using Administrative Claims”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.1 (2021-05-18). Number: 1, pp. 453–460. ISSN: 2374-3468. DOI: [10.1609/aaai.v35i1.16122](https://doi.org/10.1609/aaai.v35i1.16122) (cit. on p. 136).
- [247] C. Prins et al. “Cardiac surgery risk-stratification models”. In: *Cardiovascular Journal of Africa* 23.3 (2012-04-20), pp. 160–164. ISSN: 19951892, 16800745. DOI: [10.5830/CVJA-2011-047](https://doi.org/10.5830/CVJA-2011-047) (cit. on p. 55).
- [248] P. Pudil, J. Novovičová, and J. Kittler. “Floating search methods in feature selection”. In: *Pattern Recognition Letters* 15.11 (1994-11), pp. 1119–1125. ISSN: 01678655. DOI: [10.1016/0167-8655\(94\)90127-9](https://doi.org/10.1016/0167-8655(94)90127-9) (cit. on p. 60).
- [249] A. Qayyum et al. “Secure and Robust Machine Learning for Healthcare: A Survey”. In: *IEEE Reviews in Biomedical Engineering* 14 (2021). Conference Name: IEEE Reviews in Biomedical Engineering, pp. 156–180. ISSN: 1941-1189. DOI: [10.1109/RBME.2020.3013489](https://doi.org/10.1109/RBME.2020.3013489) (cit. on pp. 31, 33).
- [250] J. Qiu et al. “Large AI Models in Health Informatics: Applications, Challenges, and the Future”. In: *IEEE Journal of Biomedical and Health Informatics* 27.12 (2023-12), pp. 6074–6087. ISSN: 2168-2194, 2168-2208. DOI: [10.1109/JBHI.2023.3316750](https://doi.org/10.1109/JBHI.2023.3316750) (cit. on pp. 132, 133).

- [251] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1.1 (1986-03-01), pp. 81–106. ISSN: 1573-0565. DOI: [10.1007/BF00116251](https://doi.org/10.1007/BF00116251) (cit. on p. 12).
- [252] T. P. Quinn et al. "The three ghosts of medical AI: Can the black-box present deliver?" In: *Artificial Intelligence in Medicine* 124 (2022-02-01), p. 102158. ISSN: 0933-3657. DOI: [10.1016/j.artmed.2021.102158](https://doi.org/10.1016/j.artmed.2021.102158) (cit. on p. 3).
- [253] J. Le-Rademacher and X. Wang. "Time-To-Event Data: An Overview and Analysis Considerations". In: *Journal of Thoracic Oncology* 16.7 (2021-07-01), pp. 1067–1074. ISSN: 1556-0864. DOI: [10.1016/j.jtho.2021.04.004](https://doi.org/10.1016/j.jtho.2021.04.004) (cit. on p. 88).
- [254] A. Radford et al. *Improving Language Understanding by Generative Pre-Training*. 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (visited on 2024-09-11) (cit. on p. 133).
- [255] A. Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021-02-26. DOI: [10.48550/arXiv.2103.00020](https://doi.org/10.48550/arXiv.2103.00020). arXiv: [2103.00020\[cs\]](https://arxiv.org/abs/2103.00020) (cit. on p. 128).
- [256] C. Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023-09-19. DOI: [10.48550/arXiv.1910.10683](https://doi.org/10.48550/arXiv.1910.10683). arXiv: [1910.10683\[cs,stat\]](https://arxiv.org/abs/1910.10683) (cit. on p. 133).
- [257] W. Raghupathi and V. Raghupathi. "Big data analytics in healthcare: promise and potential". In: *Health Information Science and Systems* 2 (2014-02-07), p. 3. ISSN: 2047-2501. DOI: [10.1186/2047-2501-2-3](https://doi.org/10.1186/2047-2501-2-3) (cit. on pp. 3, 35, 134).
- [258] K. N. V. P. S. Rajesh and R. Dhuli. "Classification of ECG heartbeats using nonlinear decomposition methods and support vector machine". In: *Computers in biology and medicine* 87 (2017-08). Publisher: Comput Biol Med, pp. 271–284. ISSN: 1879-0534. DOI: [10.1016/J.COMPBIOMED.2017.06.006](https://doi.org/10.1016/J.COMPBIOMED.2017.06.006) (cit. on p. 116).
- [259] A. Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021-02-26. DOI: [10.48550/arXiv.2102.12092](https://doi.org/10.48550/arXiv.2102.12092). arXiv: [2102.12092\[cs\]](https://arxiv.org/abs/2102.12092) (cit. on p. 128).
- [260] L. Rasmy et al. "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: *npj Digital Medicine* 4.1 (2021-05-20), p. 86. ISSN: 2398-6352. DOI: [10.1038/s41746-021-00455-y](https://doi.org/10.1038/s41746-021-00455-y) (cit. on pp. 134, 136, 150, 151, 225).
- [261] M. N. Rastgoo et al. "Automatic driver stress level classification using multimodal deep learning". In: *Expert Systems with Applications* 138 (2019-12), p. 112793. ISSN: 09574174. DOI: [10.1016/j.eswa.2019.07.010](https://doi.org/10.1016/j.eswa.2019.07.010) (cit. on p. 114).
- [262] R. Rastogi and S. Mortaza. "Multi-label classification with Missing Labels using Label Correlation and Robust Structural Learning". In: *Knowledge-Based Systems* 229 (2021-10), p. 107336. ISSN: 09507051. DOI: [10.1016/j.knosys.2021.107336](https://doi.org/10.1016/j.knosys.2021.107336) (cit. on pp. 89, 102).

- [263] A. Rath et al. "Heart disease detection using deep learning methods from imbalanced ECG samples". In: *Biomedical Signal Processing and Control* 68 (2021-07). Publisher: Elsevier, p. 102820. ISSN: 1746-8094. DOI: [10.1016/j.bspc.2021.102820](https://doi.org/10.1016/j.bspc.2021.102820) (cit. on p. 134).
- [264] S. Reddy et al. "Evaluation framework to guide implementation of AI systems into healthcare settings". In: *BMJ Health & Care Informatics* 28.1 (2021-10), e100444. ISSN: 2632-1009. DOI: [10.1136/bmjhci-2021-100444](https://doi.org/10.1136/bmjhci-2021-100444) (cit. on p. 41).
- [265] B. Ribeiro, V. Rolla, and R. Santos. "INCOGNITUS: A Toolbox for Automated Clinical Notes Anonymization". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 187–194. DOI: [10.18653/v1/2023.eacl-demo.22](https://doi.org/10.18653/v1/2023.eacl-demo.22) (cit. on p. 166).
- [266] B. Ribeiro et al. "Layered Learning for Acute Hypotensive Episode Prediction in the ICU: An Alternative Approach". In: *2021 International Conference on e-Health and Bioengineering (EHB)*. 2021 International Conference on e-Health and Bioengineering (EHB). ISSN: 2575-5145. 2021-11, pp. 1–4. DOI: [10.1109/EHB52898.2021.9657577](https://doi.org/10.1109/EHB52898.2021.9657577) (cit. on p. 165).
- [267] B. Ribeiro et al. "Unravelling Heterogeneity: A Hybrid Machine Learning Approach to Predict Post-discharge Complications in Cardiothoracic Surgery". In: *Progress in Artificial Intelligence*. Vol. 14116. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 298–310. DOI: [10.1007/978-3-031-49011-8_24](https://doi.org/10.1007/978-3-031-49011-8_24) (cit. on p. 165).
- [268] G. F. Riley. "Administrative and claims records as sources of health care cost data". In: *Medical Care* 47.7 (2009-07), S51–55. ISSN: 1537-1948. DOI: [10.1097/MLR.0b013e31819c95aa](https://doi.org/10.1097/MLR.0b013e31819c95aa) (cit. on p. 37).
- [269] T. Rockel. *missMethods: Methods for Missing Data*. Version 0.3.0. 2022. URL: <https://cran.r-project.org/web/packages/missMethods/index.html> (cit. on p. 75).
- [270] N. Rodemund et al. "The Salzburg Intensive Care database (SICdb): an openly available critical care dataset". In: *Intensive Care Medicine* 49.6 (2023-06-01), pp. 700–702. ISSN: 1432-1238. DOI: [10.1007/s00134-023-07046-3](https://doi.org/10.1007/s00134-023-07046-3) (cit. on p. 138).
- [271] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362. ISBN: 0-262-68053-X (cit. on p. 112).

- [272] M. Rupp, O. Peter, and T. Pattipaka. “ExBEHRT: Extended Transformer for Electronic Health Records to Predict Disease Subtypes & Progressions”. In: vol. 13932. 2023, pp. 73–84. DOI: [10.1007/978-3-031-39539-0_7](https://doi.org/10.1007/978-3-031-39539-0_7). arXiv: [2303.12364\[cs\]](https://arxiv.org/abs/2303.12364) (cit. on p. 136).
- [273] S. Saadatnejad, M. Oveisi, and M. Hashemi. “LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices”. In: *IEEE Journal of Biomedical and Health Informatics* 24.2 (2020-02). Publisher: Institute of Electrical and Electronics Engineers Inc., pp. 515–523. DOI: [10.1109/JBHI.2019.2911367](https://doi.org/10.1109/JBHI.2019.2911367) (cit. on p. 113).
- [274] S. Sadasivuni et al. “Recurrent neural network circuit for automated detection of atrial fibrillation from raw ECG”. In: *Proceedings - IEEE International Symposium on Circuits and Systems* 2021-May (2021). Publisher: Institute of Electrical and Electronics Engineers Inc. DOI: [10.1109/ISCAS51556.2021.9401666](https://doi.org/10.1109/ISCAS51556.2021.9401666) (cit. on p. 113).
- [275] T. Saito and M. Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”. In: *PLOS ONE* 10.3 (2015-03-04). Ed. by G. Brock. Publisher: Public Library of Science (PLOS), e0118432. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432) (cit. on p. 25).
- [276] C. E. Sanchez et al. “Predictors and Risk Calculator of Early Unplanned Hospital Readmission Following Contemporary Self-Expanding Transcatheter Aortic Valve Replacement from the STS/ACC TVT Registry”. In: *Cardiovascular Revascularization Medicine* 21.3 (2020-03), pp. 263–270. ISSN: 15538389. DOI: [10.1016/j.carrev.2019.05.032](https://doi.org/10.1016/j.carrev.2019.05.032) (cit. on p. 48).
- [277] M. Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019-03-21. arXiv: [1801.04381\[cs\]](https://arxiv.org/abs/1801.04381) (cit. on p. 117).
- [278] R. Santos et al. “A Risk Prediction Framework to Optimize Remote Patient Monitoring Following Cardiothoracic Surgery”. In: *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*. Ed. by G. De Francisci Morales et al. Cham: Springer Nature Switzerland, 2023, pp. 366–371. ISBN: 978-3-031-43430-3. DOI: [10.1007/978-3-031-43430-3_32](https://doi.org/10.1007/978-3-031-43430-3_32) (cit. on p. 165).
- [279] R. Santos et al. “CORKI: A Correlation-Driven Imputation Method for Partial Annotation Scenarios in Multi-label Clinical Problems”. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Vol. 2136. Series Title: Communications in Computer and Information Science. Cham: Springer Nature Switzerland, 2025, pp. 3–18. DOI: [10.1007/978-3-031-74640-6_1](https://doi.org/10.1007/978-3-031-74640-6_1) (cit. on pp. 87, 165).

- [280] R. Santos et al. "Predicting post-discharge complications in cardiothoracic surgery: A clinical decision support system to optimize remote patient monitoring resources". In: *International Journal of Medical Informatics* 182 (2024-02-01), p. 105307. ISSN: 1386-5056. DOI: [10.1016/j.ijmedinf.2023.105307](https://doi.org/10.1016/j.ijmedinf.2023.105307) (cit. on pp. 54, 71, 165).
- [281] I. H. Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions". In: *SN Computer Science* 2.3 (2021-03-22), p. 160. ISSN: 2661-8907. DOI: [10.1007/s42979-021-00592-x](https://doi.org/10.1007/s42979-021-00592-x) (cit. on pp. 11, 22).
- [282] G. Savcisens et al. "life2vec: Using Sequences of Life-events to Predict Human Lives". In: *Nature Computational Science* (2023-12-18). ISSN: 2662-8457. DOI: [10.1038/s43588-023-00573-5](https://doi.org/10.1038/s43588-023-00573-5). arXiv: [2306.03009](https://arxiv.org/abs/2306.03009)[cs,stat] (cit. on p. 136).
- [283] M. Scherpf et al. "Predicting sepsis with a recurrent neural network using the MIMIC III database". In: *Computers in Biology and Medicine* 113 (2019-10-01), p. 103395. ISSN: 0010-4825. DOI: [10.1016/j.combiomed.2019.103395](https://doi.org/10.1016/j.combiomed.2019.103395) (cit. on p. 136).
- [284] P. Schober, C. Boer, and L. A. Schwarte. "Correlation Coefficients: Appropriate Use and Interpretation". In: *Anesthesia & Analgesia* 126.5 (2018-05), p. 1763. ISSN: 0003-2999. DOI: [10.1213/ANE.0000000000002864](https://doi.org/10.1213/ANE.0000000000002864) (cit. on pp. 20, 78, 95).
- [285] A. M. Sefidian and N. Daneshpour. "Estimating missing data using novel correlation maximization based methods". In: *Applied Soft Computing* 91 (2020-06-01), p. 106249. ISSN: 1568-4946. DOI: [10.1016/j.asoc.2020.106249](https://doi.org/10.1016/j.asoc.2020.106249) (cit. on pp. 72, 73, 80).
- [286] T. Sethi et al. "Interpretable artificial intelligence: Closing the adoption gap in healthcare". In: *Artificial Intelligence in Precision Health*. Ed. by D. Barh. Academic Press, 2020-01-01, pp. 3–29. ISBN: 978-0-12-817133-2. DOI: [10.1016/B978-0-12-817133-2.00001-X](https://doi.org/10.1016/B978-0-12-817133-2.00001-X) (cit. on p. 40).
- [287] E. Sezgin. "Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers". In: *Digital Health* 9 (2023-01-01). Publisher: SAGE Publications Ltd, p. 20552076231186520. ISSN: 2055-2076. DOI: [10.1177/20552076231186520](https://doi.org/10.1177/20552076231186520) (cit. on p. 34).
- [288] D. M. Shahian et al. "The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1—Background, Design Considerations, and Model Development". In: *The Annals of Thoracic Surgery* 105.5 (2018-05), pp. 1411–1418. ISSN: 00034975. DOI: [10.1016/j.athoracsur.2018.03.002](https://doi.org/10.1016/j.athoracsur.2018.03.002) (cit. on pp. 49, 54, 55, 66).
- [289] K. Shameer et al. "Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort". In: *Biocomputing 2017. Proceedings of the Pacific Symposium*. Kohala Coast, Hawaii, USA: World Scientific, 2017-01, pp. 276–287. DOI: [10.1142/9789813207813_0027](https://doi.org/10.1142/9789813207813_0027) (cit. on p. 31).

- [290] M. S. R. Shawon et al. "Patient and hospital factors associated with 30-day readmissions after coronary artery bypass graft (CABG) surgery: a systematic review and meta-analysis". In: *Journal of Cardiothoracic Surgery* 16.1 (2021-12), p. 172. ISSN: 1749-8090. DOI: [10.1186/s13019-021-01556-1](https://doi.org/10.1186/s13019-021-01556-1) (cit. on p. 48).
- [291] B. Shickel et al. "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis". In: *IEEE Journal of Biomedical and Health Informatics* 22.5 (2018-09). Conference Name: IEEE Journal of Biomedical and Health Informatics, pp. 1589–1604. ISSN: 2168-2208. DOI: [10.1109/JBHI.2017.2767063](https://doi.org/10.1109/JBHI.2017.2767063) (cit. on p. 10).
- [292] A. W. Shroyer et al. "The society of thoracic surgeons: 30-day operative mortality and morbidity risk models". In: *The Annals of Thoracic Surgery* 75.6 (2003-06), pp. 1856–1865. ISSN: 00034975. DOI: [10.1016/S0003-4975\(03\)00179-6](https://doi.org/10.1016/S0003-4975(03)00179-6) (cit. on p. 56).
- [293] S. N. Shukla and B. M. Marlin. *mTAN: Multi-Time Attention Networks for Irregularly Sampled Time Series*. 2021-06-07. arXiv: [2101.10318\[cs\]](https://arxiv.org/abs/2101.10318) (cit. on p. 156).
- [294] O. Simeone. *A Brief Introduction to Machine Learning for Engineers*. 2018-05-17. DOI: [10.48550/arXiv.1709.02840](https://doi.org/10.48550/arXiv.1709.02840). arXiv: [1709.02840\[cs,math,stat\]](https://arxiv.org/abs/1709.02840) (cit. on p. 10).
- [295] H. Sivaraks and C. A. Ratanamahatana. "Robust and accurate anomaly detection in ECG artifacts using time series motif discovery". In: *Computational and Mathematical Methods in Medicine* 2015 (2015-01). Publisher: Hindawi Limited. ISSN: 17486718. DOI: [10.1155/2015/453214](https://doi.org/10.1155/2015/453214) (cit. on p. 116).
- [296] S. Śmigiel, K. Pałczyński, and D. Ledziński. "Deep Learning Techniques in the Classification of ECG Signals Using R-Peak Detection Based on the PTB-XL Dataset". In: *Sensors* 21.24 (2021). Publisher: Multidisciplinary Digital Publishing Institute, p. 8174. DOI: [10.3390/s21248174](https://doi.org/10.3390/s21248174) (cit. on pp. 116, 126).
- [297] S. Śmigiel et al. "ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset". In: (2021). DOI: [10.3390/e23091121](https://doi.org/10.3390/e23091121) (cit. on pp. 113, 126).
- [298] J. Snoek, H. Larochelle, and R. P. Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*. 2012-08-29. arXiv: [1206.2944\[cs,stat\]](https://arxiv.org/abs/1206.2944) (cit. on p. 24).
- [299] S. Somani et al. "Deep learning and the electrocardiogram: review of the current state-of-the-art". In: *EP Europace* 23.8 (2021-08-06), pp. 1179–1191. ISSN: 1099-5129, 1532-2092. DOI: [10.1093/europace/euaa377](https://doi.org/10.1093/europace/euaa377) (cit. on p. 113).
- [300] R. Somun-Kapetanović. "Variables". In: *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1639–1640. DOI: [10.1007/978-3-642-04898-2_99](https://doi.org/10.1007/978-3-642-04898-2_99) (cit. on p. 17).

- [301] H. Song et al. "SAnD: Attend and Diagnose: Clinical Time Series Analysis Using Attention Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (2018-04-29). ISSN: 2374-3468, 2159-5399. DOI: [10.1609/aaai.v32i1.11635](https://doi.org/10.1609/aaai.v32i1.11635) (cit. on p. 136).
- [302] N. Strodthoff et al. "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL". In: *IEEE Journal of Biomedical and Health Informatics* 25.5 (2021-05). Publisher: Institute of Electrical and Electronics Engineers Inc., pp. 1519–1528. ISSN: 21682208. DOI: [10.1109/JBHI.2020.3022989](https://doi.org/10.1109/JBHI.2020.3022989) (cit. on pp. 113, 116).
- [303] A. Subbaswamy and S. Saria. "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: *Biostatistics (Oxford, England)* 21.2 (2020-04-01), pp. 345–352. ISSN: 1468-4357. DOI: [10.1093/biostatistics/kxz041](https://doi.org/10.1093/biostatistics/kxz041) (cit. on p. 3).
- [304] D. Sun, H. Zhao, and J. Sun. "Regression analysis of asynchronous longitudinal data with informative observation processes". In: *Computational Statistics & Data Analysis* 157 (2021-05-01), p. 107161. ISSN: 0167-9473. DOI: [10.1016/j.csda.2020.107161](https://doi.org/10.1016/j.csda.2020.107161) (cit. on p. 131).
- [305] T. Sun, X. He, and Z. Li. "Digital twin in healthcare: Recent updates and challenges". In: *DIGITAL HEALTH* 9 (2023-01-01). Publisher: SAGE Publications Ltd, p. 20552076221149651. ISSN: 2055-2076. DOI: [10.1177/20552076221149651](https://doi.org/10.1177/20552076221149651) (cit. on p. 33).
- [306] R. T. Sutton et al. "An overview of clinical decision support systems: benefits, risks, and strategies for success". In: *npj Digital Medicine* 3.1 (2020-02-06). Publisher: Nature Publishing Group, pp. 1–10. ISSN: 2398-6352. DOI: [10.1038/s41746-020-0021-y](https://doi.org/10.1038/s41746-020-0021-y) (cit. on p. 29).
- [307] B. Swift et al. "Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care". In: *Clinical and Translational Science* 11.5 (2018). _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12559>, pp. 450–460. ISSN: 1752-8062. DOI: [10.1111/cts.12559](https://doi.org/10.1111/cts.12559) (cit. on p. 34).
- [308] M. Syed et al. "Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review". In: *Informatics (MDPI)* 8.1 (2021-03), p. 16. ISSN: 2227-9709. DOI: [10.3390/informatics8010016](https://doi.org/10.3390/informatics8010016) (cit. on p. 135).
- [309] S. Tabarestani et al. "A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study". In: *NeuroImage* 206 (2020-02-01), p. 116317. ISSN: 1053-8119. DOI: [10.1016/j.neuroimage.2019.116317](https://doi.org/10.1016/j.neuroimage.2019.116317) (cit. on p. 73).

- [310] S. Taheri Moghadam et al. "The effects of clinical decision support system for prescribing medication on patient outcomes and physician practice performance: a systematic review and meta-analysis". In: *BMC Medical Informatics and Decision Making* 21.1 (2021-03-10), p. 98. ISSN: 1472-6947. DOI: [10.1186/s12911-020-01376-8](https://doi.org/10.1186/s12911-020-01376-8) (cit. on p. 29).
- [311] S. Tahzeeb and S. Hasan. "A Neural Network-Based Multi-Label Classifier for Protein Function Prediction". In: *Engineering, Technology & Applied Science Research* 12.1 (2022-02-12), pp. 7974–7981. ISSN: 1792-8036, 2241-4487. DOI: [10.48084/etasr.4597](https://doi.org/10.48084/etasr.4597) (cit. on p. 88).
- [312] A. Tan et al. "Weak multi-label learning with missing labels via instance granular discrimination". In: *Information Sciences* 594 (2022-05), pp. 200–216. ISSN: 00200255. DOI: [10.1016/j.ins.2022.02.011](https://doi.org/10.1016/j.ins.2022.02.011) (cit. on p. 89).
- [313] W. Tang et al. *Rethinking 1D-CNN for Time Series Classification: A Stronger Baseline*. 2020-02-23. DOI: [10.48550/arXiv.2002.10061](https://doi.org/10.48550/arXiv.2002.10061) (cit. on p. 15).
- [314] A. N. Tarekegn, M. Giacobini, and K. Michalak. "A review of methods for imbalanced multi-label classification". In: *Pattern Recognition* 118 (2021-10), p. 107965. ISSN: 00313203. DOI: [10.1016/j.patcog.2021.107965](https://doi.org/10.1016/j.patcog.2021.107965) (cit. on p. 88).
- [315] K. M. Tarwani and S. Edem. "Survey on Recurrent Neural Network in Natural Language Processing". In: *International Journal of Engineering Trends and Technology* 48.6 (2017-06-25), pp. 301–304. ISSN: 22315381. DOI: [10.14445/22315381/IJETT-V48P253](https://doi.org/10.14445/22315381/IJETT-V48P253) (cit. on p. 15).
- [316] S. Taylor et al. *Responsible AI – Key Themes, Concerns & Recommendations For European Research And Innovation*. Zenodo, 2018-07-02. DOI: [10.5281/ZENODO.1303252](https://doi.org/10.5281/ZENODO.1303252) (cit. on p. 41).
- [317] P. J. Thorat et al. "Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example*". In: *Critical Care Medicine* 49.6 (2021-06), e563. ISSN: 0090-3493. DOI: [10.1097/CCM.0000000000004916](https://doi.org/10.1097/CCM.0000000000004916) (cit. on p. 138).
- [318] S. Tipirneni and C. K. Reddy. "STraTS: Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series". In: *ACM Transactions on Knowledge Discovery from Data* 16.6 (2022-12-31), pp. 1–17. ISSN: 1556-4681, 1556-472X. DOI: [10.1145/3516367](https://doi.org/10.1145/3516367) (cit. on pp. 129, 136, 156).
- [319] E. J. Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature Medicine* 25.1 (2019-01), pp. 44–56. ISSN: 1078-8956, 1546-170X. DOI: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7) (cit. on pp. 31, 34, 39).

- [320] O. Troyanskaya et al. "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6 (2001-06-01), pp. 520–525. ISSN: 1367-4811, 1367-4803. DOI: [10.1093/bioinformatics/17.6.520](https://doi.org/10.1093/bioinformatics/17.6.520) (cit. on p. 90).
- [321] C.-F. Tsai, M.-L. Li, and W.-C. Lin. "A class center based approach for missing value imputation". In: *Knowledge-Based Systems* 151 (2018-07), pp. 124–135. ISSN: 09507051. DOI: [10.1016/j.knosys.2018.03.026](https://doi.org/10.1016/j.knosys.2018.03.026) (cit. on p. 71).
- [322] A. Ullah et al. "Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation". In: *Remote Sensing* 12.10 (2020). Publisher: Multidisciplinary Digital Publishing Institute, p. 1685. DOI: [10.3390/rs12101685](https://doi.org/10.3390/rs12101685) (cit. on p. 114).
- [323] United States Food and Drug Administration. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. 2021. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> (visited on 2024-09-03) (cit. on p. 46).
- [324] S. Uygun İlkhan et al. "How to mitigate the risks of deployment of artificial intelligence in medicine?" In: *Turkish Journal of Medical Sciences* 54.3 (2024-06-12), pp. 483–492. ISSN: 1300-0144. DOI: [10.55730/1300-0144.5814](https://doi.org/10.55730/1300-0144.5814) (cit. on p. 32).
- [325] A. Vallée. "Digital twin for healthcare systems". In: *Frontiers in Digital Health* 5 (2023-09-07). Publisher: Frontiers. ISSN: 2673-253X. DOI: [10.3389/fdgth.2023.1253050](https://doi.org/10.3389/fdgth.2023.1253050) (cit. on p. 33).
- [326] A. Vaswani et al. *Attention Is All You Need*. 2017. arXiv: [1706.03762\[cs\]](https://arxiv.org/abs/1706.03762) (cit. on pp. 16, 17, 117, 118, 132, 154).
- [327] D. Vervoort et al. "Global Cardiac Surgical Volume and Gaps: Trends, Targets, and Way Forward". In: *Annals of Thoracic Surgery Short Reports* 2.2 (2024-06-01). Publisher: Elsevier, pp. 320–324. ISSN: 2772-9931. DOI: [10.1016/j.atssr.2023.11.019](https://doi.org/10.1016/j.atssr.2023.11.019) (cit. on p. 48).
- [328] J. -. Vincent et al. "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure". In: *Intensive Care Medicine* 22.7 (1996-07-01), pp. 707–710. ISSN: 1432-1238. DOI: [10.1007/BF01709751](https://doi.org/10.1007/BF01709751) (cit. on p. 30).
- [329] S. C. Virgeniya and E. Ramaraj. "A Novel Deep Learning based Gated Recurrent Unit with Extreme Learning Machine for Electrocardiogram (ECG) Signal Recognition". In: *Biomedical Signal Processing and Control* 68 (2021), p. 102779. ISSN: 1746-8094. DOI: [10.1016/j.bspc.2021.102779](https://doi.org/10.1016/j.bspc.2021.102779) (cit. on pp. 113, 116, 126).
- [330] C. N. Vorisek et al. "Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review". In: *JMIR Medical Informatics* 10.7 (2022-07-19), e35724. ISSN: 2291-9694. DOI: [10.2196/35724](https://doi.org/10.2196/35724) (cit. on p. 37).

- [331] P. Wagner et al. "PTB-XL, a large publicly available electrocardiography dataset". In: *Scientific Data* 7.1 (2020-05-25), p. 154. ISSN: 2052-4463. DOI: [10.1038/s41597-020-0495-6](https://doi.org/10.1038/s41597-020-0495-6) (cit. on pp. 92, 111, 113, 115, 116).
- [332] Z. Wang, W. Yan, and T. Oates. *Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline*. 2016-12-14. DOI: [10.48550/arXiv.1611.06455](https://doi.org/10.48550/arXiv.1611.06455). arXiv: [1611.06455\[cs, stat\]](https://arxiv.org/abs/1611.06455) (cit. on p. 113).
- [333] J. E. Ware and C. D. Sherbourne. "The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection". In: *Medical Care* 30.6 (1992-06), pp. 473–483. ISSN: 0025-7079 (cit. on p. 52).
- [334] M. Wasimuddin et al. "Stages-Based ECG Signal Analysis From Traditional Signal Processing to Machine Learning Approaches: A Survey". In: *IEEE Access* 8 (2020), pp. 177782–177803. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2020.3026968](https://doi.org/10.1109/ACCESS.2020.3026968) (cit. on p. 112).
- [335] R. van de Water et al. *Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML*. 2024-01-18. arXiv: [2306.05109\[cs\]](https://arxiv.org/abs/2306.05109) (cit. on pp. 135, 136, 144, 148, 154, 155, 222).
- [336] K. Wei et al. "MS-Net: Sleep apnea detection in PPG using multi-scale block and shadow module one-dimensional convolutional neural network". In: *Computers in Biology and Medicine* 155 (2023), p. 106469. ISSN: 0010-4825. DOI: [10.1016/j.combiomed.2022.106469](https://doi.org/10.1016/j.combiomed.2022.106469) (cit. on p. 113).
- [337] N. G. Weiskopf and C. Weng. "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research". In: *Journal of the American Medical Informatics Association : JAMIA* 20.1 (2013), pp. 144–151. ISSN: 1067-5027. DOI: [10.1136/amiajnl-2011-000681](https://doi.org/10.1136/amiajnl-2011-000681) (cit. on p. 31).
- [338] P. Werbos. "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10 (1990-10). Conference Name: Proceedings of the IEEE, pp. 1550–1560. ISSN: 1558-2256. DOI: [10.1109/5.58337](https://doi.org/10.1109/5.58337) (cit. on p. 14).
- [339] White House Office of Science and Technology Policy. *Blueprint for an AI Bill of Rights*. 2022. URL: <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf> (visited on 2024-09-03) (cit. on p. 43).
- [340] J. Wiens and E. S. Shenoy. "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology". In: *Clinical Infectious Diseases* 66.1 (2018-01-06), pp. 149–153. ISSN: 1058-4838, 1537-6591. DOI: [10.1093/cid/cix731](https://doi.org/10.1093/cid/cix731) (cit. on p. 135).
- [341] J. Wiens et al. "Do no harm: a roadmap for responsible machine learning for health care". In: *Nature Medicine* 25.9 (2019-09), pp. 1337–1340. ISSN: 1546-170X. DOI: [10.1038/s41591-019-0548-6](https://doi.org/10.1038/s41591-019-0548-6) (cit. on pp. 26, 28, 38).

- [342] B. M. Wiesenfeld, Y. Aphinyanaphongs, and O. Nov. "AI model transferability in healthcare: a sociotechnical perspective". In: *Nature Machine Intelligence* 4.10 (2022-10). Publisher: Nature Publishing Group, pp. 807–809. ISSN: 2522-5839. DOI: [10.1038/s42256-022-00544-x](https://doi.org/10.1038/s42256-022-00544-x) (cit. on p. 162).
- [343] T. Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020-07-13. DOI: [10.48550/arXiv.1910.03771](https://doi.org/10.48550/arXiv.1910.03771). arXiv: [1910.03771\[cs\]](https://arxiv.org/abs/1910.03771) (cit. on pp. 150, 225).
- [344] M. Wornow et al. *EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models*. 2023-12-11. DOI: [10.48550/arXiv.2307.02028](https://doi.org/10.48550/arXiv.2307.02028). arXiv: [2307.02028 \[cs\]](https://arxiv.org/abs/2307.02028) (cit. on p. 136).
- [345] A. Wright and D. F. Sittig. "A Four-Phase Model of the Evolution of Clinical Decision Support Architectures". In: *International journal of medical informatics* 77.10 (2008-10), pp. 641–649. ISSN: 1386-5056. DOI: [10.1016/j.ijmedinf.2008.01.004](https://doi.org/10.1016/j.ijmedinf.2008.01.004) (cit. on p. 29).
- [346] B. Wu et al. "Multi-label Learning with Missing Labels". In: *2014 22nd International Conference on Pattern Recognition*. 2014 22nd International Conference on Pattern Recognition (ICPR). Stockholm, Sweden: IEEE, 2014-08, pp. 1964–1968. ISBN: 978-1-4799-5209-0. DOI: [10.1109/ICPR.2014.343](https://doi.org/10.1109/ICPR.2014.343) (cit. on p. 90).
- [347] Q. Xiao et al. "Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review". In: *Applied Sciences* 13.8 (2023-04-14), p. 4964. ISSN: 2076-3417. DOI: [10.3390/app13084964](https://doi.org/10.3390/app13084964) (cit. on p. 106).
- [348] R. Xiao et al. "Integrating multimodal information in machine learning for classifying acute myocardial infarction". In: *Physiological Measurement* 44.4 (2023-04-01), p. 044002. ISSN: 0967-3334, 1361-6579. DOI: [10.1088/1361-6579/acc77f](https://doi.org/10.1088/1361-6579/acc77f) (cit. on p. 114).
- [349] L. Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023-12-19. arXiv: [2312.12148\[cs\]](https://arxiv.org/abs/2312.12148) (cit. on p. 133).
- [350] L. Xu et al. "Learning Low-Rank Label Correlations for Multi-label Classification with Missing Labels". In: *2014 IEEE International Conference on Data Mining*. 2014 IEEE International Conference on Data Mining (ICDM). Shenzhen, China: IEEE, 2014-12, pp. 1067–1072. DOI: [10.1109/ICDM.2014.125](https://doi.org/10.1109/ICDM.2014.125) (cit. on p. 90).
- [351] S. Yang et al. "A Multi-View Multi-Scale Neural Network for Multi-Label ECG Classification". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 7.3 (2023-06), pp. 648–660. ISSN: 2471-285X. DOI: [10.1109/TETCI.2023.3235374](https://doi.org/10.1109/TETCI.2023.3235374) (cit. on pp. 113, 126).
- [352] O. Yildirim et al. "A new approach for arrhythmia classification using deep coded features and LSTM networks". In: *Computer Methods and Programs in Biomedicine* 176 (2019), pp. 121–133. DOI: [10.1016/j.cmpb.2019.05.004](https://doi.org/10.1016/j.cmpb.2019.05.004) (cit. on p. 113).

- [353] J. Yoon, W. R. Zame, and M. van der Schaar. “Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks”. In: *IEEE Transactions on Biomedical Engineering* 66.5 (2019-05). Conference Name: IEEE Transactions on Biomedical Engineering, pp. 1477–1490. ISSN: 1558-2531. DOI: [10.1109/TBME.2018.2874712](https://doi.org/10.1109/TBME.2018.2874712) (cit. on p. 73).
- [354] R. Yu et al. “Using a Multi-Task Recurrent Neural Network With Attention Mechanisms to Predict Hospital Mortality of Patients”. In: *IEEE Journal of Biomedical and Health Informatics* 24.2 (2020-02). Conference Name: IEEE Journal of Biomedical and Health Informatics, pp. 486–492. ISSN: 2168-2208. DOI: [10.1109/JBHI.2019.2916667](https://doi.org/10.1109/JBHI.2019.2916667) (cit. on p. 136).
- [355] J. E. Yun et al. “Comparative Effectiveness of Telemonitoring Versus Usual Care for Heart Failure: A Systematic Review and Meta-analysis”. In: *Journal of Cardiac Failure* 24.1 (2018-01), pp. 19–28. ISSN: 10719164. DOI: [10.1016/j.cardfail.2017.09.006](https://doi.org/10.1016/j.cardfail.2017.09.006) (cit. on p. 50).
- [356] X. Zeng et al. “PIC, a paediatric-specific intensive care database”. In: *Scientific Data* 7.1 (2020-01-13), p. 14. ISSN: 2052-4463. DOI: [10.1038/s41597-020-0355-4](https://doi.org/10.1038/s41597-020-0355-4) (cit. on p. 138).
- [357] D. Zha et al. *Data-centric Artificial Intelligence: A Survey*. 2023-06-11. DOI: [10.48550/arXiv.2303.10158](https://doi.org/10.48550/arXiv.2303.10158). arXiv: [2303.10158\[cs\]](https://arxiv.org/abs/2303.10158) (cit. on p. 40).
- [358] B. Zhang, H. Shi, and H. Wang. “Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach”. In: *Journal of Multidisciplinary Healthcare* 16 (null 2023-12-31). Publisher: Dove Medical Press, pp. 1779–1791. DOI: [10.2147/JMDH.S410301](https://doi.org/10.2147/JMDH.S410301) (cit. on p. 33).
- [359] C. Zhang et al. “Hybrid Noise-Oriented Multilabel Learning”. In: *IEEE Transactions on Cybernetics* 50.6 (2020-06), pp. 2837–2850. ISSN: 2168-2267, 2168-2275. DOI: [10.1109/TCYB.2019.2894985](https://doi.org/10.1109/TCYB.2019.2894985) (cit. on pp. 89, 90).
- [360] D. Zhang et al. *Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions*. 2024-06-04. arXiv: [2307.03941\[cs\]](https://arxiv.org/abs/2307.03941) (cit. on p. 44).
- [361] X. Zhang et al. *Graph-Guided Network for Irregularly Sampled Multivariate Time Series*. 2022-03-15. DOI: [10.48550/arXiv.2110.05357](https://doi.org/10.48550/arXiv.2110.05357). arXiv: [2110.05357\[cs\]](https://arxiv.org/abs/2110.05357) (cit. on p. 156).
- [362] Y. Zhang et al. *Data-Centric Foundation Models in Computational Healthcare: A Survey*. 2024-01-04. DOI: [10.48550/arXiv.2401.02458](https://doi.org/10.48550/arXiv.2401.02458). arXiv: [2401.02458\[cs\]](https://arxiv.org/abs/2401.02458) (cit. on pp. 40, 41).
- [363] L. Zhou, M. Rueda, and A. Alkhateeb. “Classification of Breast Cancer Nottingham Prognostic Index Using High-Dimensional Embedding and Residual Neural Network”. In: *Cancers* 14.4 (2022-02-13), p. 934. ISSN: 2072-6694. DOI: [10.3390/cancers14040934](https://doi.org/10.3390/cancers14040934) (cit. on p. 71).

- [364] H. Zhu et al. "A lightweight piecewise linear synthesis method for standard 12-lead ECG signals based on adaptive region segmentation". In: *PLOS ONE* 13.10 (2018-10). ISBN: 1111111111 Publisher: Public Library of Science, e0206170. ISSN: 1932-6203. DOI: [10.1371/JOURNAL.PONE.0206170](https://doi.org/10.1371/JOURNAL.PONE.0206170) (cit. on p. 116).
- [365] J. Zhu, H. Chen, and W. Ye. "Classification of Human Activities Based on Radar Signals using 1D-CNN and LSTM". In: *2020 IEEE International Symposium on Circuits and Systems (ISCAS)* (2020), pp. 1–5. DOI: [10.1109/ISCAS45731.2020.9181233](https://doi.org/10.1109/ISCAS45731.2020.9181233) (cit. on p. 113).

SUPPLEMENTARY MATERIAL ON ROBUST CARDIOTHORACIC SURGERY RISK ESTIMATION

This Appendix includes additional information on the development of different works presented and discussed in Chapter 4. We start by detailing the collected variables in the inpatient dataset from Section 4.2.2.1. From Sections 4.3, 4.4 and 4.5, we present the detailed hyperparameter ranges tested for different imputation methods and classifiers, as well as the best combinations. Additionally, flowcharts for the proposed imputation methods in Section 4.4 and extended results regarding Section 4.5 are included.

A.1 Inpatient Dataset Description

The extensive list of variables collected by the clinical team since 2008 is available in Table A.1. These were used to build the models throughout Chapter 4.

APPENDIX A. SUPPLEMENTARY MATERIAL ON ROBUST CARDIOTHORACIC SURGERY RISK ESTIMATION

Table A.1: List of variables by group in the inpatient dataset. O denotes an ordinal variable, B a binary, N a nominal, NU a numerical, I an identifier, and T a textual field.

Inpatient Dataset Variables
<i>Hospitalization</i>
Date of admission; Date of death; Date of discharge; Date of surgery
<i>Cardiac history</i>
Angina (O); Dyspnoea (O); Number of previous MIs (O); Most recent MI (O); Congestive heart failure (B)
<i>Previous interventions</i>
Previous PCI (O); Date of last PCI; Date of last cardiac surgery; Previous cardiac, vascular, or thoracic surgery (N)
<i>Pre-operative risk factors</i>
Weight (NU); Height (NU); Smoking history (O); Diabetes treatment (O); Hypercholesterolaemia (B); Hypertension (B); Renal disease (N); Last pre-operative creatinine (NU); Chronic lung disease (O); Extra-cardiac arteriopathy (N); Cerebrovascular disease (N); Carotid bruits (B); Neurological dysfunction (B); Pre-operative heart rhythm (N)
<i>Pre-operative haemodynamics & catheterization</i>
Left- or right-heart catheter (N); Date of last catheterization; Number of diseased coronary vessels (O); Left main stem disease (B); Ejection fraction category (O); Ejection fraction value (NU); PA systolic (NU); AV gradient (NU); LVEDP (NU); Mean PAWP / LA (NU)
<i>Pre-operative status & support</i>
IV inotropes (B); IV nitrates/heparin of any kind (B); Ventilated (B); Cardiogenic shock (B)
<i>Operation</i>
Operative urgency (O); Main reason for urgency (N); Number of previous heart operations (O); Procedure groups (N); Other cardiac procedures (N); Other non-cardiac procedures (N); Other operation description (T)
<i>Coronary surgery</i>
Arterial distal coronary anastomoses (NU); Venous distal coronary anastomoses (NU); Arteries used as grafts (N)
<i>Valve surgery</i>
Valve procedures (NU); Valve site (N); Stenosis (B); Insufficiency (O); Explant type (N); Native valve pathology (N); Reason to repeat valve surgery (N); Valve procedure (B); Implant type (N); Implant code (I); Implant size (NU); Valve status (O); Segments of the aorta (O); Aortic procedure (N); Cardiopulmonary bypass (N); Predominant form of myocardial protection (B); Cardioplegia: Infusion mode (B); Cardioplegia: Solution (B); Cardioplegia: Temperature (B); Cardioplegia: Timing (B); Intra-aortic balloon pump (N); Non-cardioplegia myocardial protection (N); Reason for IABP use (N); Bypass time (NU); Total circulatory arrest time (NU); Cumulative X-clamp time (NU); Duration of surgery (NU); ICU stay (NU); Intermediate stay (NU)
<i>Cardiac Surgery Morbidity Scale</i>
CNS complications (N); Renal complications (N); Respiratory complications (N); Cardiac complications (N); Detailed cardiac complications (N); Bleeding complications (N); Infective complications (N); Other complications (N); Morbidity Index (NU)
<i>Discharge details</i>
Destination on discharge (N); Patient status at discharge (B); Primary cause of death (N)
<i>Patient demographics and auto-calculations</i>
Age (NU); Gender (B); Hosp No (I); BMI (NU); BSA (NU); Additive EuroSCORE (NU); EuroSCORE results (NU); Logistic EuroSCORE (NU); Pre-operative stay (NU); Post-Operative stay (NU); Total hospital stay (NU); Status (N); Subsequent entry (NU)

A.2 Detailed Experiments on Developing Risk Models for Cardiothoracic Surgery

The optimization of prediction pipelines under Section 4.3 tested a range of hyperparameters. They are available in Table A.2.

Table A.2: Hyperparameter ranges for the risk prediction model. The optimal values for each parameter are emboldened.

Classifier	Hyperparameter Values
NB	$\text{var_smoothing} \in \{1.0 \times 10^{-4}; 2.8 \times 10^{-5}; 7.7 \times 10^{-6}; 2.2 \times 10^{-6}; 6.0 \times 10^{-7}; 1.7 \times 10^{-7}; 4.6 \times 10^{-8}; 1.3 \times 10^{-8}; 3.6 \times 10^{-9}; 1.0 \times 10^{-9}\}$
SVM	$\text{class_weight} = \text{'balanced'}$; $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$; $\text{gamma} \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
DT	$\text{class_weight} = \text{'balanced'}$; $\text{criterion} \in \{\text{'entropy'}, \text{'gini'}\}$; $\text{max_depth} \in \{3, 5, 7, 9\}$; $\text{min_samples_split} \in \{20, 40, 50, 100, 200, 500\}$; $\text{min_samples_leaf} \in \{20, 40, 50, 100, 200, 300\}$; $\text{max_features} \in \{\text{'log2'}, \text{'sqrt'}, \text{None}\}$; $\text{max_leaf_nodes} \in \{20, 40, 60, 80, 100, 200\}$
RF	$\text{class_weight} = \text{'balanced'}$; $\text{max_depth} \in \{3, 4, 5\}$; $\text{n_estimators} \in \{50, 100, 200, 500, 1000, 2000\}$; $\text{criterion} \in \{\text{'entropy'}, \text{'gini'}\}$; $\text{min_samples_split} \in \{30, 40, 50, 100, 200, 500, 1000\}$; $\text{min_samples_leaf} \in \{25, 30, 50, 200, 500\}$; $\text{max_leaf_nodes} \in \{\text{None}, 5, 10, 20\}$
LightGBM	$\text{class_weight} = \text{'balanced'}$; $\text{objective} = \text{'binary'}$; $\text{metric} = \text{'roc_auc'}$; $\text{max_depth} \in \{3, 4, 5\}$; $\text{num_leaves} \in \{4, 9, 25, 50, 100\}$; $\text{min_data_in_leaf} \in \{30, 50, 200, 500\}$; $\text{learning_rate} \in \{0.01, 0.001, 0.0001\}$; $\text{n_estimators} \in \{50, 100, 200, 500, 1000, 2000\}$; $\text{lambda_l1} \in \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$; $\text{lambda_l2} \in \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$

A.3 Complementary Material on Handling Missing Data from the Real World

The work developed under Section 4.4 proposed new correlation-based imputation methods that improve the robustness against missing data. Figures A.1a, A.1b and A.2 present flowcharts for the CWKNNI, KNNSCI and CWRI methods, respectively.

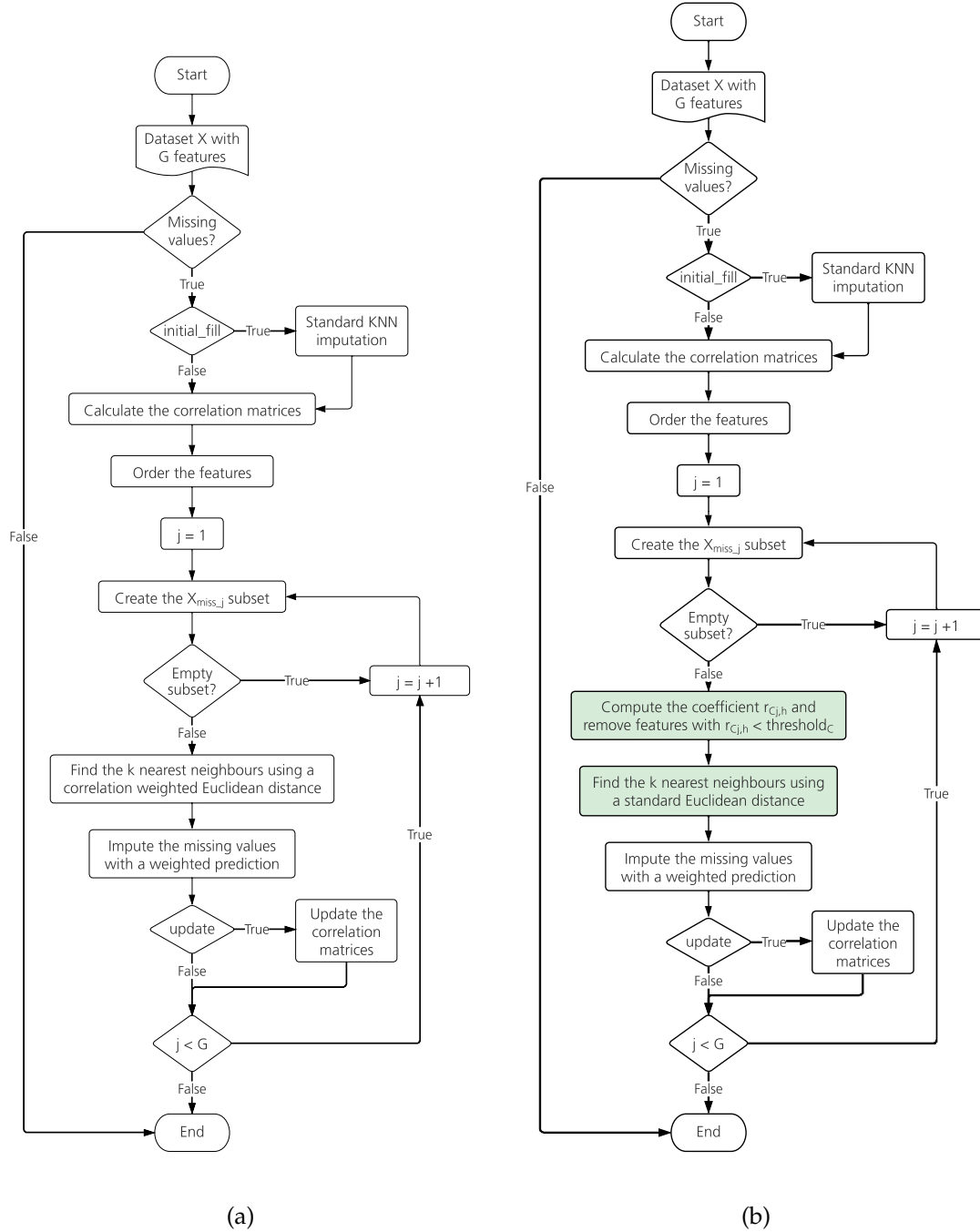


Figure A.1: Flowcharts of the (a) CWKNNI and (b) KNNSCI methods. Green boxes denote the structural differences of KNNSCI to CWKNNI.

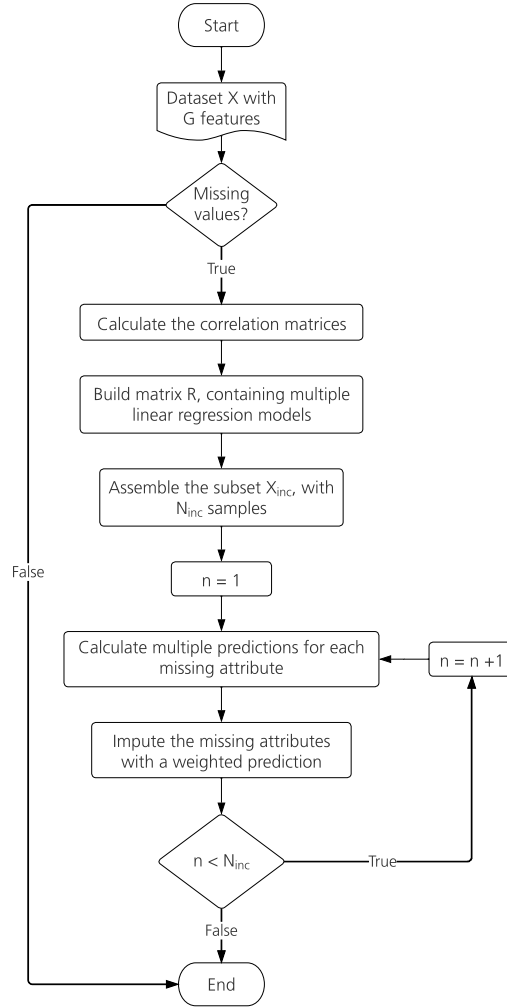


Figure A.2: Flowchart of the CWRI method.

A.4 Detailed Experiments on Handling Missing Data from the Real World

Each proposed imputation method (CWKNNI, KNNSCI and CWRI) was compared to several alternative imputation strategies from the literature. These were optimized according to the range of parameters presented in Table A.3. In addition, model performance comparisons were performed using three different ML models (RF, SVM, and NB), which were also optimized with distinct ranges for the UCI datasets and the cardiothoracic surgery dataset. The hyperparameter ranges are presented in Table A.4. The resulting best combination between model and imputation method for every dataset is shown in Table A.5.

APPENDIX A. SUPPLEMENTARY MATERIAL ON ROBUST CARDIOTHORACIC SURGERY RISK ESTIMATION

Table A.3: Tested parameter values for every imputation method within the comparative study. N.A. denotes imputation methods that do not depend on parameter specification.

Imputation Method	Parameter Values
Mean / Mode	missing_values = np.nan
Regression	N.A.
KNN	missing_values = np.nan; n_neighbors $\in \{5, 10, 15\}$; weights = 'distance'; metric = 'nan_euclidean'
CMIM	percentage $\in \{0.1, 0.5, 0.9\}$; threshold $\in \{0.1, 0.5, 0.9\}$
CoHiKNN	n_neighbors $\in \{5, 10, 15\}$
NMVI	N.A.
MICE	m = 3; maxit = 10; seed = 42; defaultMethod = c("pmm", "logreg", "polyreg", "polr")
CWKNNI	k $\in \{5, 10, 15\}$; p $\in \{0.2, 0.4, 0.6, 0.8\}$; initial_fill $\in \{\text{False}, \text{True}\}$; update $\in \{\text{False}, \text{True}\}$
KNNSCI	k $\in \{5, 10, 15\}$; p $\in \{0.2, 0.4, 0.6, 0.8\}$; initial_fill $\in \{\text{False}, \text{True}\}$; update $\in \{\text{False}, \text{True}\}$
CWRI	p $\in \{0.2, 0.4, 0.6, 0.8\}$

Table A.4: Hyperparameter ranges for every classifier and dataset source.

Classifier	Hyperparameter Values
<i>UCI Machine Learning Repository datasets</i>	
RF	max_depth $\in \{3, 5, 7, 9\}$; n_estimators $\in \{5, 10, 25, 50, 100, 200\}$; criterion $\in \{\text{'entropy'}, \text{'gini'}\}$; min_samples_split $\in \{10, 20, 50\}$; min_samples_leaf $\in \{5, 10, 25\}$
SVM	random_state = 42; class_weight = 'balanced'; C $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$; gamma $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	var_smoothing $\in \{1.0 \times 10^{-4}, 2.8 \times 10^{-5}, 7.7 \times 10^{-6}, 2.2 \times 10^{-6},$ $6.0 \times 10^{-7}, 1.7 \times 10^{-7}, 4.6 \times 10^{-8}, 1.3 \times 10^{-8}, 3.6 \times 10^{-9}, 1.0 \times 10^{-9}\}$
<i>Cardiothoracic Surgery dataset</i>	
RF	max_depth $\in \{3, 6, 9\}$; n_estimators $\in \{10, 25, 100\}$; criterion $\in \{\text{'entropy'}, \text{'gini'}\}$; min_samples_split $\in \{20, 50\}$; min_samples_leaf $\in \{10, 25\}$
SVM	random_state = 42; class_weight = 'balanced'; C $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$; gamma $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	var_smoothing $\in \{1.0 \times 10^{-4}, 2.8 \times 10^{-5}, 7.7 \times 10^{-6}, 2.2 \times 10^{-6},$ $6.0 \times 10^{-7}, 1.7 \times 10^{-7}, 4.6 \times 10^{-8}, 1.3 \times 10^{-8}, 3.6 \times 10^{-9}, 1.0 \times 10^{-9}\}$

Table A.5: Best imputation and classification hyperparameters and parameters for each dataset, in terms of the highest average AUROC.

Classifier	Hyperparameters	Imputation	Parameters
<i>Wine dataset</i>			
RF	random_state = 42; max_depth = 3; n_estimators = 200; criterion='entropy'; min_samples_split = 10; min_samples_leaf = 10	MICE	N.A.
SVM	random_state = 42; class_weight = 'balanced'; C = 1; gamma = 1	CWRI	$p = 0.8$
NB	var_smoothing = 1.0×10^{-4}	CWRI	N.A.
<i>SPECT Heart dataset</i>			
RF	random_state = 42; max_depth = 5; n_estimators = 100; criterion='entropy'; min_samples_split = 20; min_samples_leaf = 5	KNNSCI	$k = 10; p = 0.8; update = \text{True};$ $initial_fill = \text{False}$
SVM	random_state = 42; class_weight = 'balanced'; C = 0.1; gamma = 0.1	NMVI	N.A.
NB	var_smoothing = 1.0×10^{-4}	CWKNNI	$k = 15; p = 0.2; update = \text{False};$ $initial_fill = \text{True}$
<i>Statlog Heart dataset</i>			
RF	random_state = 42; max_depth = 5; n_estimators = 200; criterion='entropy'; min_samples_split = 10; min_samples_leaf = 5	CWRI	$p = 0.8$
SVM	random_state = 42; class_weight = 'balanced'; C = 100; gamma = 0.01	CWRI	$p = 0.8$
NB	var_smoothing = 1.0×10^{-4}	MICE	N.A.
<i>Cardiothoracic surgery (2008-2018) dataset</i>			
RF	random_state = 42; max_depth = 6; n_estimators = 100; criterion='gini'; min_samples_split = 20; min_samples_leaf = 25	KNNSCI	$k = 15; p = 0.2; update = \text{False};$ $initial_fill = \text{False}$
SVM	random_state = 42; class_weight = 'balanced'; C = 0.1; gamma = 0.1	MICE	N.A.
NB	var_smoothing = 1.7×10^{-7}	MICE	N.A.
<i>Cardiothoracic surgery (2019) dataset</i>			
RF	random_state = 42; max_depth = 3; n_estimators = 100; criterion='gini'; min_samples_split = 20; min_samples_leaf = 10	KNNSCI	$k = 10; p = 0.6; update = \text{True};$ $initial_fill = \text{False}$
SVM	random_state = 42; class_weight = 'balanced'; C = 10; gamma = 1	Regression	N.A.
NB	var_smoothing = 1.7×10^{-7}	MICE	N.A.

A.5 Detailed Experiments on Providing Longitudinal Risk Predictions in Partial Annotation Scenarios

In Section 4.5, we presented a novel method for missing label imputation in multi-label scenarios, as is the case of longitudinal predictions in cardiothoracic surgery. CORKI is based on previous developments and uses correlation to improve the imputation process. To evaluate the performance of CORKI in different use cases, we ran a hyperparameter tuning phase to optimize the models for each task. Table A.6 details the range of tested hyperparameters.

Table A.6: Tested hyperparameter values for every classifier and dataset source tested in CORKI.

Classifier	Hyperparameter Values
RF	$\text{max_depth} \in \{3, 4, 5\}; \text{n_estimators} \in \{50, 100, 200\};$
	$\text{criterion} \in \{\text{'entropy'}, \text{'gini'}\}; \text{min_samples_leaf} \in \{25, 50, 100, 200\};$
	$\text{max_leaf_nodes} \in \{\text{None}, 5, 10, 20\}; \text{max_features} \in \{\text{None}, \text{'log2'}, \text{'sqrt'}\};$
	$\text{class_weight} = \text{'balanced'}$
DT	$\text{max_depth} \in \{3, 5, 7, 9\}; \text{criterion} \in \{\text{'entropy'}, \text{'gini'}\};$
	$\text{min_samples_leaf} \in \{25, 50, 100, 200\}; \text{max_leaf_nodes} \in \{\text{None}, 5, 10, 20\};$
	$\text{max_features} \in \{\text{None}, \text{'auto'}, \text{'log2'}, \text{'sqrt'}\}; \text{class_weight} = \text{'balanced'}$
MLP	$\text{hidden_layer_sizes} \in \{2, 10, 50, 100, 200\}; \text{solver} \in \{\text{'sgd'}, \text{'adam'}\};$
	$\text{activation} \in \{\text{'identity'}, \text{'logistic'}, \text{'tanh'}, \text{'relu'}\};$
	$\text{learning_rate_init} \in \{0.01, 0.001, 0.0001\}; \text{batch_size} \in \{16, 32, 128\}$
KNN	$\text{n_neighbors} \in \{3, 5, 6, 10, 15, 25, 50\}; \text{weights} \in \{\text{'uniform'}, \text{'distance'}\};$
	$\text{metric} \in \{\text{'minkowski'}, \text{'cityblock'}, \text{'cosine'}, \text{'euclidean'}, \text{'manhattan'}\}$

A.6 Extended Results on Providing Longitudinal Risk Predictions in Partial Annotation Scenarios

Experiments to evaluate CORKI were conducted in the cardiothoracic surgery use case and a multi-label ECG classification task. Tables A.7 and A.8 present extended results of the experiments conducted and thoroughly described in the respective section.

Table A.7: Complete results of ECG classification for the test set in the CPSC, PTB-XL, and G12EC datasets. Results are presented as percentage (%) in terms of AUROC, sensitivity, specificity, and F1-score, as the mean and standard deviation of all labels. CORKI P. refers to the disease prevalence weighting. MLP was generally selected as the best model, except in cases marked with *, where KNN was selected.

	MR	Strategy	K	AUROC	Sensitivity	Specificity	F1-Score
CPSC	3	Original	-	92.9 $\pm$ 3.9	83.6 $\pm$ 8.5	88.3 $\pm$ 6.7	61.2 $\pm$ 16.7
		Drop	-	92.9 $\pm$ 3.9	83.0 $\pm$ 7.6	89.4 $\pm$ 4.6	61.2 $\pm$ 16.4
		CORKI	50	93.0 $\pm$ 3.9	85.3 $\pm$ 6.3	87.0 $\pm$ 8.9	59.1 $\pm$ 21.1
	5	CORKI P.	5	92.5 $\pm$ 4.0	85.3 $\pm$ 6.9	85.9 $\pm$ 10.0	57.8 $\pm$ 22.7
		Drop	-	92.6 $\pm$ 4.1	82.6 $\pm$ 10.4	90.0 $\pm$ 5.5	60.6 $\pm$ 21.7
		CORKI	50	93.0 $\pm$ 3.8	85.3 $\pm$ 4.8	87.7 $\pm$ 7.8	59.5 $\pm$ 20.4
	10	CORKI P.	5	92.4 $\pm$ 4.0	84.5 $\pm$ 6.4	87.1 $\pm$ 6.9	57.9 $\pm$ 22.0
		Drop	-	92.0 $\pm$ 4.0	81.9 $\pm$ 9.1	89.3 $\pm$ 4.2	59.3 $\pm$ 15.6
		CORKI	20	92.7 $\pm$ 3.9	83.1 $\pm$ 8.1	89.1 $\pm$ 6.8	62.0 $\pm$ 15.7
	25	CORKI P.	5	92.3 $\pm$ 3.8	86.0 $\pm$ 7.0	85.8 $\pm$ 8.3	57.4 $\pm$ 18.9
		Drop	-	81.0 $\pm$ 9.2	57.4 $\pm$ 30.9	79.1 $\pm$ 21.0	38.5 $\pm$ 23.6
		CORKI	50	92.1 $\pm$ 4.4	81.6 $\pm$ 10.2	89.2 $\pm$ 4.6	59.6 $\pm$ 15.9
	50	CORKI P.	50	89.4 $\pm$ 5.2	87.4 $\pm$ 7.5	70.6 $\pm$ 6.9	40.4 $\pm$ 18.8
		Drop	-	-	-	-	-
		CORKI	50	87.4 $\pm$ 6.6	73.8 $\pm$ 13.3	89.3 $\pm$ 4.3	55.7 $\pm$ 16.2
		CORKI P.	5	87.4 $\pm$ 5.9	84.0 $\pm$ 11.5	71.7 $\pm$ 11.1	40.4 $\pm$ 18.2

APPENDIX A. SUPPLEMENTARY MATERIAL ON ROBUST CARDIOTHORACIC SURGERY RISK ESTIMATION

Continuation of Table A.7						
MR	Strategy	K	AUROC	Sensitivity	Specificity	F1-Score
PTB-XL	Original	-	91.6 $\pm$ 8.6	86.2 $\pm$ 5.9	82.5 $\pm$ 16.5	45.1 $\pm$ 28.9
	Drop	-	90.6 $\pm$ 11.1	79.4 $\pm$ 13.7	86.8 $\pm$ 9.3	46.1 $\pm$ 28.9
	CORKI	50	90.3 $\pm$ 12.4	80.8 $\pm$ 13.1	83.0 $\pm$ 16.8	45.6 $\pm$ 28.9
	CORKI P.	5	90.3 $\pm$ 11.5	84.3 $\pm$ 8.6	80.5 $\pm$ 20.3	44.1 $\pm$ 28.9
	Drop	-	90.7 $\pm$ 10.1	81.5 $\pm$ 13.0	85.5 $\pm$ 8.2	44.0 $\pm$ 29.3
	CORKI	5	90.5 $\pm$ 12.2	84.0 $\pm$ 8.4	83.0 $\pm$ 16.2	45.8 $\pm$ 28.9
	CORKI P.	5	90.3 $\pm$ 12.1	85.3 $\pm$ 9.0	81.7 $\pm$ 15.3	44.2 $\pm$ 28.4
	Drop	-	85.8 $\pm$ 16.2	72.2 $\pm$ 28.6	87.8 $\pm$ 8.7	43.8 $\pm$ 29.4
	CORKI	5	90.8 $\pm$ 10.8	84.2 $\pm$ 8.4	83.6 $\pm$ 13.4	44.6 $\pm$ 28.2
	CORKI P.	5	90.3 $\pm$ 11.2	87.2 $\pm$ 8.5	80.0 $\pm$ 15.9	42.4 $\pm$ 28.8
	Drop	-	-	-	-	-
	CORKI	5	90.2 $\pm$ 11.9	84.3 $\pm$ 9.1	81.7 $\pm$ 16.4	43.9 $\pm$ 27.4
	CORKI P.	5	89.8 $\pm$ 10.2	89.4 $\pm$ 10.6	75.5 $\pm$ 14.2	37.6 $\pm$ 27.7
	Drop	-	-	-	-	-
	CORKI	50	90.0 $\pm$ 10.1	87.5 $\pm$ 8.5	78.9 $\pm$ 20.8	43.1 $\pm$ 29.1
	CORKI P.	5	87.7 $\pm$ 10.9	90.0 $\pm$ 9.6	72.1 $\pm$ 12.0	33.6 $\pm$ 28.7
G12EC	Original	-	86.5 $\pm$ 13.4	73.6 $\pm$ 20.7	85.7 $\pm$ 6.2	44.8 $\pm$ 30.5
	Drop	-	85.7 $\pm$ 12.1	70.6 $\pm$ 23.7	82.8 $\pm$ 13.4	44.6 $\pm$ 32.1
	CORKI	20	86.2 $\pm$ 13.9	71.2 $\pm$ 20.2	83.1 $\pm$ 11.7	45.1 $\pm$ 31.8
	CORKI P.	5	85.4 $\pm$ 14.8	70.1 $\pm$ 21.6	84.4 $\pm$ 9.6	44.5 $\pm$ 30.6
	Drop	-	85.1 $\pm$ 13.8	69.5 $\pm$ 23.5	81.7 $\pm$ 16.8	44.9 $\pm$ 31.7
	CORKI	5	85.9 $\pm$ 14.6	72.8 $\pm$ 19.4	85.4 $\pm$ 7.0	45.4 $\pm$ 30.9
	CORKI P.	5	84.8 $\pm$ 15.9	70.7 $\pm$ 24.1	83.5 $\pm$ 8.8	44.1 $\pm$ 31.4
	Drop	-	82.4 $\pm$ 16.1	70.2 $\pm$ 22.8	82.1 $\pm$ 7.8	42.1 $\pm$ 30.1
	CORKI	50	86.2 $\pm$ 13.7	74.6 $\pm$ 22.4	84.5 $\pm$ 8.3	45.6 $\pm$ 32.0
	CORKI P.	5	84.4 $\pm$ 15.0	72.4 $\pm$ 20.0	82.1 $\pm$ 9.6	43.1 $\pm$ 31.1
	Drop	-	65.7 $\pm$ 22.9	31.4 $\pm$ 25.2	92.1 $\pm$ 7.8	30.1 $\pm$ 24.8
	CORKI	20	84.8 $\pm$ 16.5	67.4 $\pm$ 26.5	86.9 $\pm$ 6.4	45.3 $\pm$ 29.8
	CORKI P.	50	82.7 $\pm$ 13.2	78.5 $\pm$ 18.1	73.5 $\pm$ 11.1	37.0 $\pm$ 28.1
	Drop	-	-	-	-	-
	CORKI	20	83.7 $\pm$ 15.9	71.1 $\pm$ 23.1	84.4 $\pm$ 4.9	43.8 $\pm$ 29.6
	CORKI P.	5	81.4 $\pm$ 14.0	76.2 $\pm$ 17.9	71.0 $\pm$ 11.6	35.2 $\pm$ 27.5

A.6. EXTENDED RESULTS ON PROVIDING LONGITUDINAL RISK PREDICTIONS IN PARTIAL ANNOTATION SCENARIOS

Table A.8: Complete classification results of longitudinal risk estimation in the cardio-thoracic surgery test sets. Results are presented as percentage (%) in terms of AUROC, sensitivity, specificity and F1-score, as the mean and standard deviation of all labels. CORKI P. refers to the disease prevalence weighting. MLP was always selected as the best model. * corresponds to the same outcome in all three tested K.

	Strategy	K	AUROC	Sensitivity	Specificity	F1-Score
2008-2018	Drop	-	57.3 $\pm$ 2.6	43.4 $\pm$ 20.8	64.2 $\pm$ 10.1	5.9 $\pm$ 1.3
	Domain	-	67.9 $\pm$ 4.7	53.4 $\pm$ 7.6	70.1 $\pm$ 2.2	9.0 $\pm$ 2.8
	CORKI	*	66.3 $\pm$ 4.1	51.7 $\pm$ 5.9	68.4 $\pm$ 1.8	8.3 $\pm$ 2.6
	CORKI P.	50	67.9 $\pm$ 5.9	57.1 $\pm$ 14.0	66.3 $\pm$ 5.0	8.6 $\pm$ 2.9
2019	Drop	-	55.6 $\pm$ 3.5	40.5 $\pm$ 13.2	64.8 $\pm$ 13.4	6.5 $\pm$ 0.7
	Domain	-	60.5 $\pm$ 3.2	44.6 $\pm$ 3.5	66.8 $\pm$ 3.4	7.5 $\pm$ 1.2
	CORKI	*	58.6 $\pm$ 1.6	47.3 $\pm$ 4.7	69.4 $\pm$ 2.0	8.5 $\pm$ 1.2
	CORKI P.	50	57.0 $\pm$ 3.7	37.9 $\pm$ 9.3	67.4 $\pm$ 6.4	6.5 $\pm$ 1.3

SUPPLEMENTARY MATERIAL ON ROBUST ECG DIAGNOSIS

This Appendix includes complementary experimental details and extended results of Chapter 5, specifically related to the work of Section 5.4.

B.1 Detailed Experiments on Improving Cardiovascular Disease Diagnoses using Electrocardiograms

The training of 1D and 2D unimodal strategies for ECG classification, followed by the multimodal fusion approaches, was preceded by a hyperparameter tuning phase. The range of hyperparameters considered is listed in Table B.1, while the optimal values are described in Table B.2.

Table B.1: Hyperparameter ranges for the unimodal and multimodal fusion models.

Parameter	1D Models		2D Models	Multimodal Fusion Models		
	RNN-based	CNN-RNN-based		Late	Early	Joint
Hidden layers	{2; 3}	3 (CNN), 1 (RNN)	N.A. [†]	1	3	3
Hidden units	{128; 256}	{128; 256}	N.A. [†]	{64; 128; 256; 512}	{128; 256; 512}	{128; 256; 512}
Filters	N.A.	N.A.	{8; 16; 32; 64}	N.A.	N.A.	N.A.
Dropout	{0; 0.3; 0.5}	{0; 0.3}	{0; 0.3; 0.5}	0	0	0
Optimizer	Adam	Adam	{Adam; SGD}	Adam	Adam	Adam
Learning rate	0.005	0.005	{0.1; 0.01; 0.001}	{0.1; 0.01; 0.001}	{0.1; 0.01; 0.001}	{0.1; 0.01; 0.001}
Batch size	256 (128) [*]	256 (128) [*]	{32; 64; 256}	256	128	64

^{*} The batch size was reduced to 128 in cases of limited memory.

[†] The numbers of layers and units depend on the proposed settings for the trained architectures.

Table B.2: Optimal hyperparameters for each unimodal and multimodal fusion model.

Model	Selected Parameters
<i>1D Models</i>	
GRU	hidden layers = 3; hidden units = 128; dropout = 0
GRU Attention	hidden layers = 3; hidden units = 128; dropout = 0
LSTM	hidden layers = 2; hidden units = 256; dropout = 0
BiGRU	hidden layers = 2; hidden units = 128; dropout = 0.5
BiLSTM	hidden layers = 2; hidden units = 128; dropout = 0.5
CNN-GRU	hidden units = 256; dropout = 0.3
CNN-LSTM	hidden units = 128; dropout = 0.3
<i>2D Models</i>	
AlexNet	filters = 16; batch size = 256; learning rate = 0.01; optimizer = Adam; dropout = 0
ResNet	filters = 16; batch size = 128; learning rate = 0.01 optimizer = SGD; dropout = 0
VGGNet	filters = 16; batch size = 128; learning rate = 0.1 optimizer = Adam; dropout = 0.3
MobileNetV2	filters = 32; batch size = 16; learning rate = 0.1 optimizer = Adam; dropout = 0
AlexNetAtt	filters = 8; batch size = 16; learning rate = 0.01 optimizer = Adam; dropout = 0
<i>Multimodal Fusion Models</i>	
Late	final hidden units = 128; learning rate = 0.001
Late Attention	final hidden units = 128; learning rate = 0.01
Early	final hidden units = 128 (from the fifth convolutional layer); learning rate = 0.1
Early Attention	final hidden units = 256 (from the fourth convolutional layer); learning rate = 0.01
Joint	final hidden units = 256 (from the last convolutional layer); learning rate = 0.01
Joint Attention	final hidden units = 128 (from the first dense layer); learning rate = 0.001

B.2 Extended Results on Improving Cardiovascular Disease Diagnoses using Electrocardiograms

The 1D, 2D, and multimodal fusion models were trained in a multi-label setting. Table B.3 presents extended results per disease for 1D models, while Table B.4 includes results for 2D models. Finally, Table B.5 consists of the complete results for the multimodal models.

Table B.3: Extended 1D ECG classification results obtained with RNN and CNN-RNN models, in percentage (%). Spec., Sens., and G-Mean denote specificity, sensitivity, and the geometric mean of these values, respectively. The best results are highlighted in bold.

1D Strategy	Metric	MI	STTC	CD	HYP	AD	Total	G-Mean
GRU	Sens.	84.81	83.37	78.71	72.24	77.11	79.67	80.35
	Spec.	78.82	83.96	86.31	71.74	87.20	81.04	
LSTM	Sens.	73.60	82.79	82.12	76.81	68.24	75.42	77.72
	Spec.	85.90	84.39	78.68	65.63	90.80	80.09	
BiGRU	Sens.	79.20	82.41	80.12	76.43	76.89	78.95	80.00
	Spec.	82.24	83.78	87.93	67.05	88.16	81.07	
BiLSTM	Sens.	78.84	84.89	78.71	71.10	75.68	78.18	79.51
	Spec.	82.17	80.79	87.33	70.05	87.12	80.87	
CNN-GRU	Sens.	72.88	84.13	67.47	66.54	69.55	72.33	75.28
	Spec.	78.88	75.30	83.78	70.63	86.16	78.35	
CNN-LSTM	Sens.	71.61	88.15	68.67	63.50	73.38	74.04	75.53
	Spec.	77.70	71.95	84.98	68.84	84.88	77.06	
GRU Attention	Sens.	83.72	80.88	76.71	78.33	69.11	76.55	78.33
	Spec.	80.50	86.16	85.89	63.53	89.44	80.15	

Table B.4: Extended 2D ECG classification results obtained with CNN models, in percentage (%). Spec., Sens., and G-Mean denote specificity, sensitivity, and the geometric mean of these two values, respectively. The best results are highlighted in bold.

2D Strategy	Metric	MI	STTC	CD	HYP	AD	Total	G-Mean
AlexNet	Sens.	64.56	82.98	73.09	68.06	76.34	73.85	75.11
	Spec.	74.22	76.95	77.96	73.63	80.48	76.38	
ResNet50	Sens.	68.54	78.20	68.67	65.40	60.13	67.31	71.75
	Spec.	73.73	79.39	77.36	69.37	85.84	76.48	
VGGNet	Sens.	67.81	84.51	67.07	66.16	79.63	74.62	75.19
	Spec.	72.17	75.49	80.00	73.21	79.04	75.77	
MobileNetV2	Sens.	57.68	63.86	57.83	32.32	83.57	65.05	73.03
	Spec.	75.40	82.50	85.71	90.95	71.20	81.98	
AlexNetAtt	Sens.	60.94	74.76	59.84	58.56	82.69	70.36	75.95
	Spec.	77.45	83.66	88.11	82.68	76.40	81.98	

Table B.5: Extended multimodal fusion ECG classification results, in percentage (%). Spec., Sens., and G-Mean denote specificity, sensitivity, and the geometric mean of these two values, respectively. The best results are highlighted in bold.

2D Strategy	Metric	MI	STTC	CD	HYP	AD	Total	G-Mean
Late Fusion	Sens.	75.59	88.91	82.73	77.19	76.34	79.82	79.42
	Spec.	86.27	77.56	83.18	64.11	88.72	79.02	
Late Fusion Attention	Sens.	79.75	84.13	76.31	72.24	86.64	81.53	74.35
	Spec.	67.83	75.30	55.56	64.11	79.84	67.80	
Early Fusion	Sens.	80.47	85.47	79.12	68.44	74.15	77.93	78.80
	Spec.	76.40	79.15	84.32	73.11	88.40	79.68	
Early Fusion Attention	Sens.	73.06	77.82	78.31	71.48	79.41	76.87	76.20
	Spec.	68.39	80.12	82.70	66.11	83.52	75.54	
Joint Fusion	Sens.	70.71	83.37	71.29	72.62	71.52	73.67	76.56
	Spec.	81.18	78.11	87.03	67.21	88.16	79.55	
Joint Fusion Attention	Sens.	77.94	88.15	77.71	66.16	65.61	74.62	77.16
	Spec.	81.06	78.41	84.74	69.11	89.60	79.79	

SUPPLEMENTARY MATERIAL ON ROBUST CONTINUOUS PATIENT MONITORING

This Appendix includes additional details related to the proposed ICU-BERT model in Chapter 6, specifically related to the work of Section 6.4. We present details related to the data utilized from MIMIC-IV and YAIB datasets and the specific phenotypes targeted by DuETT and replicated in this work. Furthermore, we include tested hyperparameter ranges and extended results.

C.1 Additional Data and Task Details on Developing Transformer Models for Critical Care

ICU-BERT was pre-trained in the MIMIC-IV database, which includes rich data on hospital and ICU stays in the Beth Israel Deaconess Medical Center. The database tables from which relevant medical registries were retrieved are detailed in Table C.1. Furthermore, we validated ICU-BERT in additional databases processed with the YAIB framework. The resulting datasets include 52 features listed in Table C.2. Finally, among the real-world tasks used to validate ICU-BERT, phenotyping is a multi-label classification task, whose 25 phenotypes are listed in Table C.3.

APPENDIX C. SUPPLEMENTARY MATERIAL ON ROBUST CONTINUOUS PATIENT MONITORING

Table C.1: Tables extracted from MIMIC-IV database and used in ICU-BERT.

MIMIC-IV Database Sources		
<i>hosp</i> Module		
admissions	diagnoses_icd	d_icd_diagnoses
drgcodes	labevents	d_labitems
microbiologyevents	omr	patients
prescriptions	procedures_icd	d_icd_procedures
services	transfers	
<i>icu</i> Module		
chartevents	d_items	datetimeevents
icustays	ingredientevents	inputevents
outputevents	procedureevents	
<i>derived</i> Tables		
antibiotic	apsiii	bg
blood_differential	cardiac_marker	charlson
chemistry	coagulation	complete_blood_count
crrt	enzyme	gcs
height	icp	inflammation
invasive_line	kdigo_stages	lods
meld	neuroblock	norepinephrine_equivalent_dose
oasis	oxygen_delivery	sapsii
sepsis3	sirs	sofa
urine_output	vasoactive_agent	ventilation
ventilator_setting	vitalsign	weight_durations

Table C.2: Clinical concepts used by *ricu* to produce YAIB fine-tuning benchmarks. Retrieved from [335].

Feature	<i>ricu</i> code	Unit
<i>Static Variables</i>		
Age at hospital admission	age	Years
Female sex	sex	N.A.
Patient height	height	cm
Patient weight	weight	kg
<i>Dynamic Variables</i>		
Blood pressure (systolic)	sbp	mmHg
Blood pressure (diastolic)	dbp	mmHg
Heart rate	hr	beats/minute
Mean arterial pressure	map	mmHg

C.1. ADDITIONAL DATA AND TASK DETAILS ON DEVELOPING
TRANSFORMER MODELS FOR CRITICAL CARE

Continuation of Table C.2

Feature	<i>ricu</i> code	Unit
Oxygen saturation	o2sat	%
Respiratory rate	resp	breaths/minute
Temperature	temp	°C
Albumin	alb	g/dL
Alkaline phosphatase	alp	IU/L
Alanine aminotransferase	alt	IU/L
Aspartate aminotransferase	ast	IU/L
Base excess	be	mmol/L
Bicarbonate	bicar	mmol/L
Bilirubin (total)	bili	mg/dL
Bilirubin (direct)	bili_dir	mg/dL
Band form neutrophils	bnd	%
Blood urea nitrogen	bun	mg/dL
Calcium	ca	mg/dL
Calcium ionized	cai	mmol/L
Creatinine	crea	mg/dL
Creatinine kinase	ck	IU/L
Creatinine kinase MB	ckmb	ng/mL
Chloride	cl	mmol/L
CO ² partial pressure	pco2	mmHg
C-reactive protein	crp	mg/L
Fibrinogen	fgn	mg/dL
Glucose	glu	mg/dL
Haemoglobin	hgb	g/dL
International normalised ratio (INR)	inr_pt	N.A.
Lactate	lact	mmol/L
Lymphocytes	lymph	%
Mean cell haemoglobin	mch	pg
Mean corpuscular haemoglobin concentration	mchc	%
Mean corpuscular volume	mcv	fL
Methaemoglobin	methb	%
Magnesium	mg	mg/dL
Neutrophils	neut	%
O ² partial pressure	po2	mmHg
Partial thromboplastin time	ptt	sec
pH of blood	ph	N.A.
Phosphate	phos	mg/dL

APPENDIX C. SUPPLEMENTARY MATERIAL ON ROBUST CONTINUOUS
PATIENT MONITORING

Continuation of Table C.2

Feature	ricu code	Unit
Platelets	plt	1,000 / μ L
Potassium	k	mmol/L
Sodium	na	mmol/L
Troponin T	tnt	ng/mL
White blood cells	wbc	1,000 / μ L
Fraction of inspired oxygen	fio2	%
Urine output	urine	mL

Table C.3: ICU phenotypes used in the phenotyping fine-tuning task. Retrieved from [127].

Phenotype	Type
Acute and unspecified renal failure	acute
Acute cerebrovascular disease	acute
Acute myocardial infarction	acute
Cardiac dysrhythmias	mixed
Chronic kidney disease	chronic
Chronic obstructive pulmonary disease	chronic
Complications of surgical/medical care	acute
Conduction disorders	mixed
Congestive heart failure; nonhypertensive	mixed
Coronary atherosclerosis and related	chronic
Diabetes mellitus with complications	mixed
Diabetes mellitus without complication	chronic
Disorders of lipid metabolism	chronic
Essential hypertension	chronic
Fluid and electrolyte disorders	acute
Gastrointestinal hemorrhage	acute
Hypertension with complications	chronic
Other liver diseases	mixed
Other lower respiratory disease	acute
Other upper respiratory disease	acute
Pleurisy; pneumothorax; pulmonary collapse	acute
Pneumonia	acute
Respiratory failure; insufficiency; arrest	acute
Septicemia (except in labor)	acute
Shock	acute

C.2 Detailed Experiments on Developing Transformer Models for Critical Care

Developing ICU-BERT included a hyperparameter tuning phase before pre-training and fine-tuning to ensure the best combination would be used. The tested hyperparameters and the selection of the best values are listed in Table C.4 for pre-training and Table C.5 for fine-tuning.

Table C.4: Hyperparameter ranges and selected values for ICU-BERT pre-training.

Parameter	Values Tested	Optimal Value
<i>Multi-Token Embeddings</i>		
Pre-trained model	BioBERT	BioBERT [181]
Hidden dimension	N.A.	768
Embeddings aggregation	{concatenation; summation}	summation
<i>BERT Block</i>		
Maximum sequence length	{512; 1024}	512
No. hidden layers	N.A.*	6
No. attention heads	N.A.*	6
Hidden layers dimension	N.A.*	192
Intermediate layer size	N.A.*	64
Attention dropout	N.A. [†]	0.1
Hidden layers dropout	N.A. [†]	0.1
<i>Optimization</i>		
Optimizer	N.A.*	AdamW
Learning rate	N.A.*	$5e - 5$
Weight decay	0	0
Scheduler	N.A. [†]	linear
Warm-up	40 epochs	40 epochs
Patient batch sorting	{random; one patient sample per batch}	one patient sample per batch
Effective batch size (per GPU)	512	512
MLVM proportion [‡]	{(50%, 25%, 25%); (20%, 40%, 40%)}	(50%, 25%, 25%)
Loss weight α	{3, 5}	3
Loss weight β	1	1

* Parameter value follows Med-BERT architecture [260].

[†] Parameter value follows original BERT architecture [76] defined by the Transformers library [343].

[‡] The proportions tested refer to masking: (feature and value, only feature, only value).

Table C.5: Hyperparameter ranges and selected values for ICU-BERT fine-tuning.

Parameter	Values Tested	Optimal Value
<i>Classification Head</i>		
Head source	{[CLS] token; sequence}	[CLS] token
Sequence merge strategy	{mean; max; mean-max; bilstm; lstm; bigru; gru}	N.A. [†]
Layers selection	{last; all; last four; second last} [*]	last
Layers merge strategy	{concatenate; sum; mean}	N.A. [†]
Dropout	{0.0; 0.1; 0.2; 0.3; 0.4; 0.5}	0.5
Unfreezed layers BERT block	{0; 1; 2; 3; 4; 5; 6}	5
<i>Optimization</i>		
Optimizer	{Adam; AdamW; SGD}	AdamW
Learning rate	$[1e^{-5}, 1e^{-1}]$, log-scale	$1e^{-3}$
Learning rate BERT block	$[1e^{-6}, 1e^{-1}]$, log-scale	$1e^{-3}$
Weight decay	0	0
Scheduler	{linear; cosine; polynomial; constant}	linear
Warm-up	{0.0; 0.1; 0.2; 0.3; 0.4; 0.5}	0.0
Focal loss	{yes; no}	no
Patient batch sorting	one patient sample per batch	one patient sample per batch
Effective batch size (per GPU)	512	512

^{*} Layers definition follows tests performed in the original BERT architecture [76].

[†] This parameter is irrelevant in combination with the previous parameter.

C.3 Extended Results on Developing Transformer Models for Critical Care

ICU-BERT was fine-tuned in multiple real-world tasks. Figure C.1 presents the few-shot learning results for the hospital mortality and phenotyping classification tasks replicated based on DuETT.

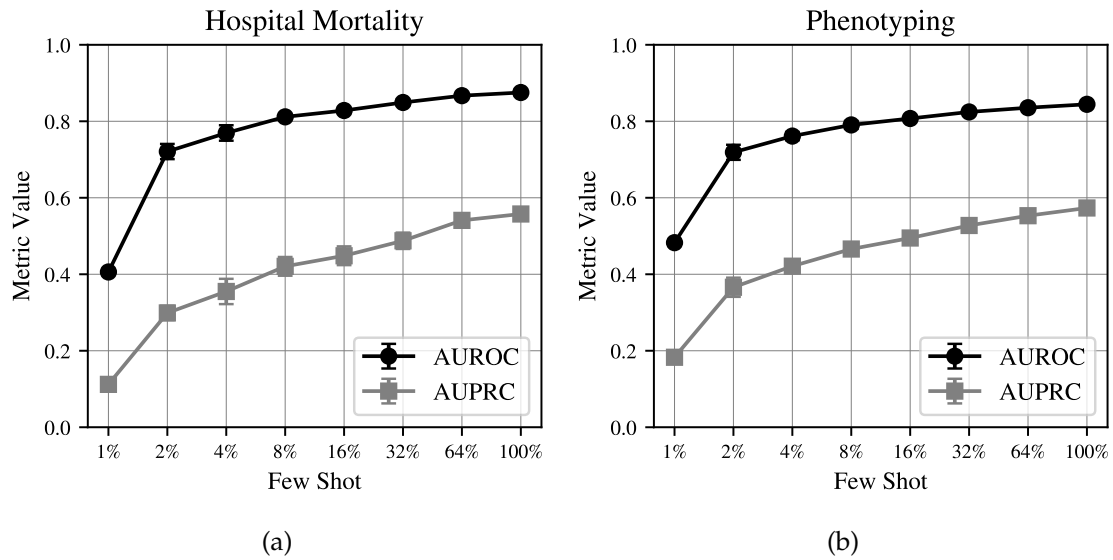


Figure C.1: Few-shot results for the (a) Hospital Mortality and (b) Phenotyping classification tasks in terms of AUROC and AUPRC (mean and standard deviation over 5-fold CV) in the MIMIC-IV dataset using 1% to 100% of all patients in the train set.



2024 Improving the Robustness of Multimodal AI with Asynchronous and Missing Inputs Ricardo Santos

