RIČARDS MARCINKEVIČS

EXPLORE, SUPPORT, AND INTERACT: SCALING INTERPRETABLE AND EXPLAINABLE MACHINE LEARNING UP TO REALITIES OF BIOMEDICAL DATA

DISS. ETH NO. 30466

DISS. ETH NO. 30466

EXPLORE, SUPPORT, AND INTERACT: SCALING INTERPRETABLE AND EXPLAINABLE MACHINE LEARNING UP TO REALITIES OF BIOMEDICAL DATA

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

RIČARDS MARCINKEVIČS
MSc ETH Statistics, ETH Zurich

born on 28.12.1995

accepted on the recommendation of

Prof. Dr. Julia E. Vogt, examiner
Prof. Dr. Michael Brudno, co-examiner
Prof. Dr. Rajesh Ranganath, co-examiner

2024

Ričards Marcinkevičs: *Explore, Support, and Interact: Scaling Interpretable and Explainable Machine Learning up to Realities of Biomedical Data*, © 2024

Staņislavam Marcinkevičam

ABSTRACT

Performant machine learning models are becoming increasingly complex and large. Due to their black-box design, they often have limited utility in exploratory data analysis and evoke little trust in non-expert users. Interpretable and explainable machine learning research emerges from application domains where, for technical or social reasons, interpreting or explaining the model's predictions or parameters is deemed necessary. In practice, interpretability and explainability are attained by (i) constructing models understandable to users by design and (ii) developing techniques to help explain already-trained black-box models.

This thesis develops interpretable and explainable machine learning models and methods tailored to applications in biomedical and healthcare data analysis. The challenges posed by this domain require nontrivial solutions and deserve special treatment. In particular, we consider practical use cases with high-dimensional and unstructured data types, diverse application scenarios, and different stakeholder groups, which all dictate special design considerations.

We demonstrate that, beyond social and ethical value, interpretability and explainability help in (i) performing exploratory data analysis, (ii) supporting medical professionals' decisions, (iii) facilitating interaction with users, and (iv) debugging the model. Our contributions are structured in two parts, tackling distinct research questions from the perspective of biomedical and healthcare applications. Firstly, we explore how to develop and incorporate inductive biases to render neural network models interpretable. Secondly, we study how to leverage explanation methods to interact with and edit already-trained black-box models.

This work spans several model and method families, including interpretable neural network architectures, prototype- and concept-based models, and attribution methods. Our techniques are motivated by classic biomedical and healthcare problems, such as time series, survival, and medical image analysis. In addition to new model and method development, we concentrate on empirical comparison, providing proof-of-concept results on real-world biomedical benchmarks.

Thus, the primary contribution of this thesis is the development of interpretable models and explanation methods with a principled treatment of specific biomedical and healthcare data types to solve application- and user-grounded problems. Through concrete use cases, we show that interpretability and explainability are context- and user-specific and, therefore, must be studied in conjunction with their application domain. We hope that our methodological and empirical contributions pave the way for future application- and user-driven interpretable and explainable machine learning research.

ZUSAMMENFASSUNG

Die Aufgaben im maschinellen Lernen werden zunehmend komplexer und größer. Aufgrund des Black-Box-Designs vieler Modelle haben sie oft nur begrenzten Nutzen in der explorativen Datenanalyse und erwecken wenig Vertrauen bei Nicht-Experten. Die Forschung im Bereich von interpretierbaren und erklärbaren Methoden des maschinellen Lernens entsteht aus Anwendungsbereichen, in denen es aus technischen oder sozialen Gründen notwendig ist, die Vorhersagen oder Parameter des Modells interpretieren oder erklären zu können. In der Praxis wird Interpretierbarkeit und Erklärbarkeit erreicht durch (i) die Entwicklung von Modellen, die von Anwendern von vornherein verstanden werden können, und (ii) die Entwicklung von Techniken, die helfen, bereits trainierte Black-Box-Modelle im Nachhinein zu erklären.

Diese Arbeit entwickelt interpretierbare und erklärbare Methoden des maschinellen Lernens, die auf Anwendungen in der biomedizinischen und Datenanalyse zugeschnitten sind. Die Herausforderungen, die dieses Gebiet mit sich bringt, erfordern neue Lösungen und verdienen besondere Aufmerksamkeit. Insbesondere betrachten wir praktische Anwendungsfälle mit hochdimensionalen und unstrukturierten Datentypen, vielfältigen Anwendungsszenarien und unterschiedlichen Interessengruppen, die alle besondere Anforderungen benötigen.

Wir zeigen, dass Interpretierbarkeit und Erklärbarkeit über den sozialen und ethischen Wert hinaus helfen kann bei (i) der Durchführung explorativer Datenanalysen, (ii) der Unterstützung von Entscheidungen medizinischer Fachkräfte, (iii) der Erleichterung der Interaktion mit Nutzern und (iv) dem Debuggen des Modells. Unsere Beiträge sind in zwei Teile gegliedert, die unterschiedliche Forschungsfragen behandeln. Erstens untersuchen wir, wie induktiven Verzerrung integriert werden können, um neuronale Netzwerkmodelle interpretierbar zu machen. Zweitens untersuchen wir, wie erklärbare Methoden genutzt werden können, um mit bereits trainierten Black-Box-Modellen zu interagieren und diese zu bearbeiten.

Diese Arbeit umfasst mehrere unterschiedliche Modell- und Methodenfamilien: interpretierbarer neuronaler Netzwerkarchitekturen, prototypbasierte und konzeptbasierter Modelle sowie Attributionsmethoden. Unsere Techniken sind motiviert durch klassische biomedizinische Probleme wie

Zeitreihen-, Überlebens- und Bildanalyse. Neben der Entwicklung neuer Modelle und Methoden konzentrieren wir uns auch auf den empirischen Vergleich und liefern Proof-of-Concept-Ergebnisse auf realen biomedizinischen Benchmarks.

Somit liegt der primäre Beitrag dieser Arbeit in der Entwicklung interpretierbarer und erklärbarer Methoden mit einer systematischen Behandlung spezifischer biomedizinischer Datentypen, um anwendungs- und nutzerbezogene Probleme zu lösen. Durch konkrete Anwendungsfälle zeigen wir, dass Interpretierbarkeit und Erklärbarkeit kontext- und nutzerspezifisch sind und daher im Zusammenhang mit ihrem Anwendungsbereich untersucht werden müssen. Wir hoffen, dass unsere methodischen und empirischen Beiträge den Weg ebnen für zukünftige anwendungs- und nutzergetriebene Forschung im Bereich des interpretierbaren und erklärbaren maschinellen Lernen.

ACKNOWLEDGEMENTS

First and foremost, I am sincerely grateful to my academic supervisor, Prof. Dr. Julia E. Vogt, for the opportunity to pursue this doctorate and for her mentorship, patience, and openness to new ideas. Prof. Vogt has always encouraged and inspired me to seek more, scientifically and professionally. I am proud and happy to have been part of the Medical Data Science lab, whose healthy and friendly environment is another of Prof. Vogt's achievements.

I would also like to thank Prof. Dr. Fanny Yang for agreeing to be my second advisor and for the opportunity to occasionally join the SML group's meetings. I am very grateful to the co-examiners, Prof. Dr. Michael Brudno and Prof. Dr. Rajesh Ranganath, for agreeing to review my dissertation.

The research contributions presented in this thesis would have been impossible without the coauthorship of my colleagues: Laura Manduchi, Dr. Ece Ozkan, Kieran Chin-Cheong, Sonia Laguna, and Moritz Vandenhirtz. In general, I am thankful to all the present and past members of the MDS group for their company along this journey. Special thanks go to my senior colleagues, Dr. Thomas Sutter and Dr. Imant Daunhawer, for setting a great example of work ethic and scientific integrity and for their authentic support and our morning and lunchtime conversations. My gratitude goes to Laura Manduchi for her positive mindset and for the many scientifically enriching projects on which I have had the pleasure of collaborating with her. I also greatly appreciate the coauthorship of Dr. Ece Ozkan and her long-term involvement in my research. Of course, my daily academic and professional life was made easier and carefree by the MDS lab's software engineers, Kieran Chin-Cheong and Andrea Agostini, and administrative assistants, Petra Lüthi, Patricia Kilchhofer, and Jacqueline Hirschi.

Beyond the MDS lab, it was an enriching experience to be part of the Institute for Machine Learning. I am thankful for all the serendipitous research conversations and topics, especially at the doctoral lunch seminars. I am very grateful to Dr. Djordje Miladinovic and Prof. Dr. Joachim M. Buhamnn, who, during my master's, gave me the first opportunity to experience the academic world and supported my applications to doctoral programmes. During my time at the Institute, in addition to research, I was lucky to tutor and advise many master's and bachelor's students: meeting and working with younger and brighter minds was a truly humbling experience.

To me, the most exciting and attractive aspect of my work has been its interdisciplinary nature. I want to thank all my collaborators outside the Department of Computer Science and ETH Zurich. Many long-term and productive collaborations have helped me develop and appreciate an application-driven perspective on machine learning research. I express my special gratitude to Dr. Patricia Reis Wolfertstetter from the University Children's Hospital Regensburg for the opportunity to explore machine learning applications for pediatric surgery and for her reliable and resourceful coauthorship. I am also proud to have worked with Dr. Thomas Brusnchwiler and Dr. Michal Muszynski from IBM Research Europe on the bio- and physiological signal analysis.

I gratefully acknowledge the institutions and organisations which supported my doctorate, particularly the Department of Computer Science at ETH Zurich, which, after all these years, feels like home. I thank the Swiss National Science Foundation for research funding. For the technical and computational support, my gratitude goes to the IT Service Group of the Department of Computer Science and the Leonhard Med Support team.

Last but not least, I am immeasurably grateful to my close and loved ones. I thank my parents, Staņislavs and Viktorija, my brother, Aleksandrs, and the extended family for their unconditional love and support. My friends, Kirils, Arturs, David, Danni, and Flo, scattered all over Europe, made my life that much more cheerful and fun. To my spouse, Alena, I am indebted for the ears and heart she has generously lent me and her inexhaustible patience along this journey.

CONTENTS

1 INTRODUCTION 1
  1.1 Scope and Contributions 3
  1.2 Publications 7
2 IN PURSUIT OF INTERPRETABILITY AND EXPLAINABILITY 9
  2.1 What's in a Name? 9
  2.2 Preliminaries and Notation 11
  2.3 Interpretable Machine Learning 11
    2.3.1 Rule-based Models and Scoring Systems 12
    2.3.2 Sparsity-inducing Regularisation 14
    2.3.3 Generalised Linear, Generalised Additive, and Varying-Coefficient Models 16
    2.3.4 Interpretable Neural Networks 19
  2.4 Explainable Machine Learning 23
    2.4.1 Attribution Methods 23
    2.4.2 Surrogate Models 25
    2.4.3 Concept-based Explanations 27
    2.4.4 Prototype- and Case-based Explanations 28
    2.4.5 Counterfactual Explanations 29
  2.5 On Applications to Healthcare 30
  2.6 Summary 32

I ANTE HOC INTERPRETABLE MODELS

3 NONLINEAR TIME SERIES STRUCTURE LEARNING 37
  3.1 Background 38
    3.1.1 Granger Causality 38
    3.1.2 Self-explaining Neural Networks 40
  3.2 Generalised Vector Autoregressive Models 41
    3.2.1 Inferring Granger Causality 45
    3.2.2 Further Remarks 47
  3.3 Experiments and Results 48
    3.3.1 Baseline Methods 48
    3.3.2 Benchmarking Datasets 49
    3.3.3 Evaluation Metrics 51
    3.3.4 Structure Learning 52
    3.3.5 Effect Sign Inference 54

3.3.6 Ablation Analysis of the Loss Function 55
3.4 Discussion 56
3.4.1 Potential Applications 59
3.4.2 Limitations 62
3.4.3 Future Work 63
3.5 Summary 64
4 PROTOTYPE-BASED EXPLANATIONS FOR DEEP SURVIVAL ANALYSIS 65
4.1 Background 66
4.1.1 Survival Analysis and Clustering 66
4.1.2 Variational Autoencoders 70
4.1.3 Mixture Models 72
4.2 Variational Deep Survival Clustering 73
4.2.1 Generative Model 75
4.2.2 Survival Model 76
4.2.3 Evidence Lower Bound and Optimisation 77
4.2.4 Further Remarks 81
4.3 Experimental Setup 82
4.3.1 Datasets 82
4.3.2 Baselines and Ablations 86
4.3.3 Evaluation Metrics 87
4.4 Results 88
4.4.1 Clustering 88
4.4.2 Time-to-event Prediction 91
4.4.3 Non-small Cell Lung Cancer and Computed Tomography 93
4.5 Discussion 96
4.5.1 Limitations 99
4.5.2 Future Work 100
4.6 Summary 100
5 CONCEPT-BASED MODELS IN THE WILD 103
5.1 Background 104
5.1.1 Concept-based Models 104
5.1.2 Multiview and Multimodal Learning 105
5.1.3 Machine Learning for Appendicitis 106
5.2 Multiview Concept Bottleneck Models 107
5.2.1 Model Architecture 107
5.2.2 Loss Function and Optimisation 109
5.2.3 Extension to Unobserved Concepts 110

5.3 Experimental Setup 114
5.3.1 Benchmarking Datasets 114
5.3.2 Pediatric Appendicitis Dataset 115
5.3.3 Baselines and Ablations 117
5.3.4 Evaluation Metrics 118
5.4 Results 118
5.4.1 Proof of Concept on Synthetic Data 118
5.4.2 The Role of Adversarial Regulariser 120
5.4.3 Application to Pediatric Appendicitis 121
5.5 Discussion 126
5.5.1 Limitations 129
5.5.2 Future Work 130
5.6 Summary 130

II POST HOC METHODS

6 BEYOND CONCEPT BOTTLENECKS 137
6.1 Background 138
6.1.1 Related Work 138
6.2 Concept-based Interventions on Black-box Models 139
6.2.1 Assessing Intervention Effectiveness 141
6.2.2 Fine-tuning for Intervenability 143
6.3 Experimental Setup 145
6.3.1 Datasets 145
6.3.2 Baselines and Ablations 147
6.3.3 Evaluation Metrics 149
6.4 Results 149
6.4.1 Results on Synthetic Data 149
6.4.2 Results on Natural Images 152
6.4.3 Application to Deep Chest X-ray Classifiers 153
6.5 Discussion 154
6.5.1 Limitations 157
6.5.2 Future Work 158
6.6 Summary 159
7 INTERPLAY BETWEEN EXPLANATION AND FAIRNESS 161
7.1 Background 162
7.1.1 Algorithmic Fairness 163
7.1.2 Biases in Deep Chest X-ray Classifiers 165
7.1.3 Pruning and Individual Units in Neural Networks 166

7.2 Pruning and Fine-tuning for Debiasing Neural Networks 166
    7.2.1 Differentiable Classification Parity Proxies 167
    7.2.2 Pruning for Fairness 169
    7.2.3 Fine-tuning for Fairness 171
7.3 Experimental Setup 173
    7.3.1 Datasets 173
    7.3.2 Classification Models and Debiasing Methods 175
    7.3.3 Evaluation Metrics 175
7.4 Results 176
    7.4.1 Benchmarking Results 176
    7.4.2 Debiasing Deep Chest X-ray Classifiers 178
7.5 Discussion 180
    7.5.1 Limitations 182
    7.5.2 Future Work 183
7.6 Summary 184
8 CONCLUDING REMARKS 185
8.1 Summary of Contributions 185
8.2 Limitations and Outlook 188

A SUPPLEMENTARY: NONLINEAR TIME SERIES STRUCTURE LEARNING 191
A.1 Network Architectures 191
B SUPPLEMENTARY: PROTOTYPE-BASED EXPLANATIONS FOR DEEP SURVIVAL ANALYSIS 193
B.1 Variational Autoencoders: ELBO Derivation 193
B.2 Synthetic Data Generation 193
B.3 Network Architectures 195
C SUPPLEMENTARY: CONCEPT-BASED MODELS IN THE WILD 199
C.1 Synthetic Data Generation 199
C.2 Multiview Animals with Attributes 200
C.3 Pediatric Appendicitis Dataset 200
C.4 Network Architectures 201
C.5 Further Results 203
C.6 Online Prediction Tool 205
D SUPPLEMENTARY: BEYOND CONCEPT BOTTLENECKS 207
D.1 Intervention Strategies 207
D.2 Synthetic Data Generation 208

D.3 Network Architectures 210
E SUPPLEMENTARY: INTERPLAY BETWEEN EXPLANATION AND FAIRNESS 211
E.1 Decision Boundary Covariance 211
E.2 Network Architectures 213

BIBLIOGRAPHY 215

NOTATION

Below, we outline the notation and terminology used throughout the thesis. Additional relevant and specific notation will generally be explained within corresponding chapters. When necessary, we will remark on the shorthand and overridden notation.

FREQUENTLY USED SYMBOLS

$\mathcal{S}, \mathscr{S}$	a set
$\mathscr{D}$	a dataset
N	dataset size
p	feature space dimensionality
a, b, α, β	a scalar value
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\beta}$	a vector
$\boldsymbol{x}$	a feature vector
$[\boldsymbol{x}, \boldsymbol{z}]$	concatenation
i, j, k, l	indices
x_j	the j-th component of a vector
$\boldsymbol{x}_{\mathcal{S}}$	a vector of components $[x_i : i \in \mathcal{S}]$, where $\mathcal{S} \subset \mathbb{N}$
$\boldsymbol{x}_i$	the feature vector of the i-th data point
$x_{i,j}$	the j-th component of the i-th feature vector
y	a response variable
$\hat{y}$	an estimate
$\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{\Phi}$	a matrix
$\boldsymbol{X}_{i,:}$	the i-th row of a matrix
$\boldsymbol{X}_{:,j}$	the j-th column of a matrix
$\boldsymbol{X}_{i,j}$	the component (i, j) of a matrix
t	time
$\boldsymbol{x}_t$	a multivariate time series at time t
$\{\boldsymbol{x}_t\}_t$	a multivariate time series
x_t^j	the j-th variable in a multivariate time series at time t

$f(\cdot), g(\cdot), h(\cdot)$	a function
$f(\cdot)_j$	the j-th output of a multivariate function
$\boldsymbol{\beta}$	a coefficient vector
$\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}$	function parameters
$\frac{\partial f}{\partial \boldsymbol{x}}$	the partial derivative of f w.r.t. $\boldsymbol{x}$
$\boldsymbol{J}_{\boldsymbol{x}}^{f}$	the Jacobian matrix of f w.r.t. $\boldsymbol{x}$
$\ell(\cdot, \cdot)$	a loss function
$\Omega(\cdot)$	a regulariser
$\mathbf{1}_{\{\cdot\}}$	an indicator function
$\lVert\cdot\rVert_p$	the ℓ_p-norm
$\lfloor\cdot\rfloor$	the floor function
$\exp\{\cdot\}$	the exponential function
$\mathrm{softplus}(\cdot)$	the softplus activation function
$\mathrm{ReLU}(\cdot)$	the rectified linear unit activation function
$\mathrm{sigmoid}(\cdot)$	the sigmoid activation function
$\Gamma(\cdot)$	the gamma function
$d(\cdot, \cdot)$	a distance function
$p(\cdot)$	a probability distribution
$p(\cdot\|\cdot)$	a conditional probability distribution
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	a multivariate Gaussian distribution with the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$	independence between random variables $\boldsymbol{x}$ and $\boldsymbol{y}$
$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{z}$	conditional independence between random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ given $\boldsymbol{z}$
$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$	covariance between random variables $\boldsymbol{x}$ and $\boldsymbol{y}$
$D_{\mathrm{KL}}(\cdot\,\Vert\,\cdot)$	the Kullback–Leibler divergence
$\mathbb{E}_{p(\cdot)}[\cdot]$	the expected value w.r.t. a probability distribution $p(\cdot)$

ABBREVIATIONS

ACC	accuracy
AFT	accelerated failure time
ANOVA	analysis of variance
ARI	the adjusted Rand index

AS	Alvarado score
AUPR	the area under precision-recall curve
AUROC	the area under the receiver operating characteristic
AwA	the Animals with Attributes 2 dataset
BA	balanced accuracy
CAL	calibration slope
CAV	concept activation vector
CBM	concept bottleneck model
CE	cross-entropy
CNN	convolutional neural network
CT	computed tomography
CV	cross-validation
ELBO	evidence lower bound
EOD	equal opportunity difference
fMRI	functional magnetic resonance imaging
GAM	generalised additive model
GC	Granger causality
GLM	generalised linear model
GMM	Gaussian mixture model
ICU	intensive care unit
IID	independent and identically distributed
KM	Kaplan–Meier
LSTM	long short-term memory
LVM	latent variable model
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MIMIC	Medical Information Mart for Intensive Care
ML	machine learning
MLP	multilayer perceptron
MNIST	the Modified National Institute of Standards and Technology database
MSE	mean squared error

MT	multitask
NMI	normalised mutual information
NN	neural network
NSCLC	non-small cell lung cancer
PAS	pediatric appendicitis score
PET	positron emission tomography
PH	proportional hazards
PR	precision-recall
RF	random forest
ReLU	rectified linear unit
SENN	self-explaining neural network
SGVB	stochastic gradient variational Bayes
SPD	statistical parity difference
SPINN	sparse-input neural network
SRU	statistical recurrent unit
t-SNE	t-distributed stochastic neighbour embedding
TPR	true positive rate
US	ultrasonography
VAE	variational autoencoder
VAR	vector autoregression
VCM	varying-coefficient model
VLM	vision-language model
XAI	explainable artificial intelligence

1

INTRODUCTION

At the time of writing, machine learning (ML) models are claimed to have achieved near- or super-human performance at many conventional benchmarking tasks [1], such as character [2] and natural image [3] recognition. With the widespread use of opaque models, e. g. ensembles [4], [5], neural networks [6], [7], and, more recently, foundation models [8], [9], which to the untrained eye are *black-box* systems, the need for *interpretable* and *explainable* ML becomes apparent. If we seek to harness these rich model classes beyond simple prediction or generation, for instance, to produce novel or actionable insights into the relations present in the data or the model itself [10], the brute-force black-box approach has limited utility.

For the purposes of the introduction, let us provide an informal, working description of interpretability and explainability [11], which will be treated in depth in Chapter 2. *Interpretable* ML refers to models understandable to a human expert or user *by design*, i. e. before (*ante hoc*) training, validation, or introspection. Such models are also referred to as *glass* or *grey* boxes. In contrast, *explainable* ML concerns itself with developing explanation techniques that help understand black-box models after (*post hoc*) they have been trained.

While, in many cases, interpretability or explainability is not necessary, beyond their social and ethical value [12], models and methods that follow these principles may help attain a few practical goals. Among them are the use cases considered in the current thesis: (i) performing **scientific discovery and exploratory data analysis** (Chapters 3–4), (ii) building **user trust and supporting decisions** (Chapter 5), (iii) facilitating **interaction with human users** (Chapter 6), and (iv) helping **model debugging** (Chapter 7). Below, we comment on these directions, citing concrete examples from the general literature.

In particular, an interpretable model or an explanation method can lead to hypotheses about causal relationships among the variables observed [13], [14] or help *explore* and understand such links better [15]. For example, neural-network-based approaches have been used to discover unknown physics [16], [17] and test long-standing hypotheses in semiotics [18].

Moreover, performant models and methods tailored towards a concrete end-user group can garner trust and become instrumental in decision support. For instance, sparse predictive models conforming to the format of risk scores and checklists [19]–[21], ubiquitously utilised by medical professionals in the clinical setting, have better prospects for acceptance and usability in practice than a naïve black box, assuming both have comparable predictive performance.

Another important characteristic of some model and method families is their capability to be interacted with, for example, by manipulating internal parameters or representations [22]. *Human–model interaction* [23] may have special utility in decision support systems where a user works collaboratively with an ML algorithm. By editing the model's internals, a human can better understand the system's behaviour [24] and selectively correct its mistakes, improving the predictive performance.

Lastly, a classic use case of interpretability and explainability is model debugging [13], [15], [25]: interpretations and explanations can help design ML models that are more robust and generalisable and expose undesirable reliance on sensitive information. For instance, interpretable ML can reveal spurious associations [26] and discrimination against disadvantaged groups [27] within predictive models.

Much of the contemporary methods research addressing the goals listed above has exclusively concentrated on the classification tasks and simple natural image benchmarks [28]. In contrast, this thesis will study application strategies for **biomedical and healthcare data**, which give rise to nontrivial technical challenges. Such data originate from experiments and observational studies in biology and medicine [29], typical examples being computational imaging, electronic health records, and physiological signals. Notably, many problems from this domain are high stakes and, therefore, require a greater degree of accountability and transparency. Interpretability and explainability can, in part, address these needs and ease the entry of ML methods into research and clinical practice.

One major challenge in such applications is unique and rich input and output structures that cannot always be accommodated by a naïve prediction problem. For example, consider time series [30] and medical image [31] analysis, where individual observations are not independent and the feature space may be high-dimensional, comprising multiple views or even modalities [32], [33]. Such data types and tasks require specialised interpretable model classes and explanation techniques.

Another characteristic of the biomedical and healthcare domain partially addressed within this thesis is the diversity of contexts and stakeholder groups [34]. Medical researchers, clinical professionals, patients, and applied data scientists may all raise distinct questions about the data and models they face. Thus, interpretability and explainability become highly context-dependent because the data type, task, and stakeholder should all factor into the methods' design choices.

Motivated by the practical goals and challenges described so far, this doctoral thesis will address the following broad research questions through the lens of biomedical and healthcare applications.

Question 1. *How can we develop and incorporate inductive biases in neural-network-based models to render them interpretable?*

Question 2. *How can we leverage post hoc explanation methods to interact with and edit neural-network-based models?*

Specifically, throughout this manuscript, we will introduce *novel* models and methods related to Questions 1 and 2. The methodological chapters of the thesis (Chapters 3–7) are grouped into two parts (Parts I and II), each tackling the respective research question.

1.1 SCOPE AND CONTRIBUTIONS

Having outlined the general directions and questions of our research, we now pinpoint the scope and contributions of the thesis. Individual chapters comprising this manuscript may be roughly characterised w.r.t. three essential aspects: (i) model or method family, (ii) technical problem or goal addressed, and (iii) data type. These characteristics are reflected schematically in Figure 1.1, which provides a visual roadmap through the chapters. Note that Chapter 2 is not part of this figure since it does not introduce a new model or method but rather lays down the technical background on interpretable and explainable ML.

As mentioned, Part I of this thesis will treat *ante hoc* interpretable models with applications to time series, medical image, and survival analysis. The models described by Chapters 3–5 are utilised to learn nonlinear dependencies in multivariate time series (*structure learning*), discover subgroups in high-dimensional survival data, and interpret multiview medical images via clinically relevant findings. Chapters 6 and 7 from Part II leverage *post hoc* explanation methods to interact with and debias black-box models. Both chapters will consider applications to medical image classifiers.

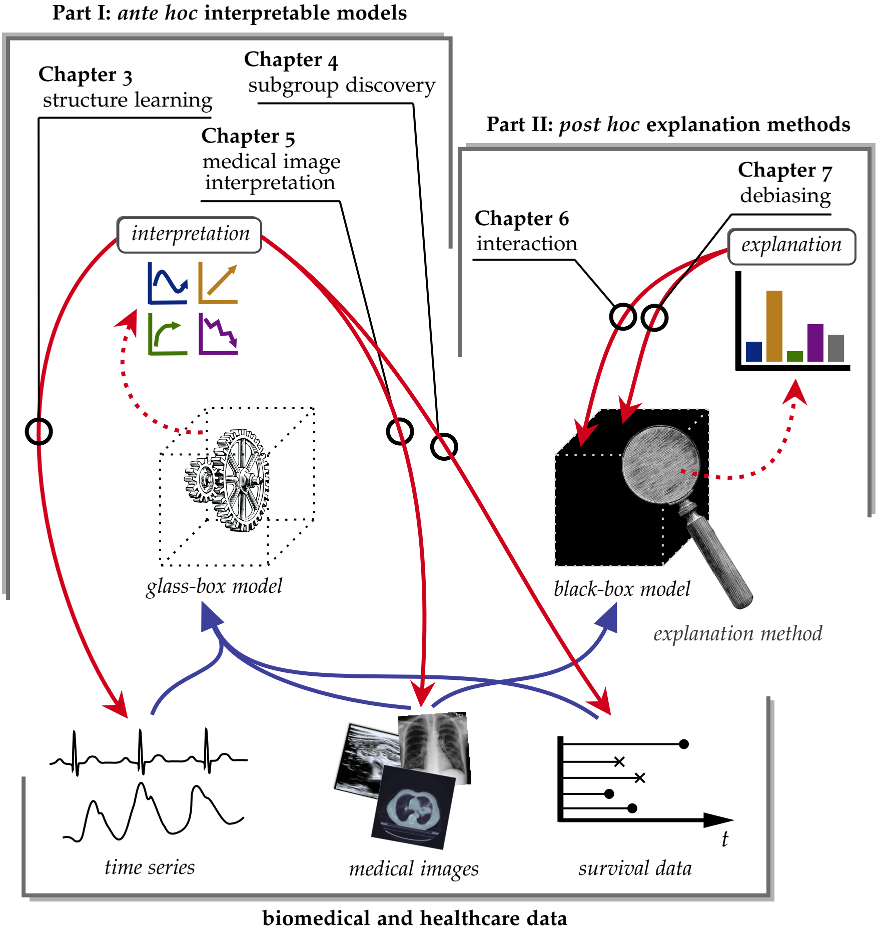


FIGURE 1.1: Schematic roadmap of the thesis. The manuscript consists of two parts: on intrinsically interpretable, i. e. *glass-box*, models (Part I) and explanation techniques for black-box approaches (Part II) in application to biomedical and healthcare data. Every chapter tackles a different technical **problem** on a specific **data type**. Note that, in this thesis, *ante hoc* interpretable models are leveraged to gain new insights about the data, whereas *post hoc* explanation methods are utilised to edit and interact with the model.

Chapter		*Ante/Post Hoc?*	Model/Method Family	Data Type	Goal	Degree of Supervision
3	*Nonlinear Time Series Structure Learning*	*ante hoc*	self-explaining neural networks	time series	structure learning	self-supervised
4	*Prototype-based Explanations for Deep Survival Analysis*	*ante hoc*	prototype-based	survival data, medical images	survival analysis, subgroup discovery	weakly supervised
5	*Concept-based Models in the Wild*	*ante hoc*	concept-based	multiview medical images	medical image interpretation, multiview classification	strongly supervised
6	*Beyond Concept Bottlenecks*	*post hoc*	concept-based	medical images	human–model interaction	strongly supervised
7	*Interplay between Explanation and Fairness*	*post hoc*	attribution	medical images	algorithmic fairness	strongly supervised

TABLE 1.1: Thesis structure. Individual chapters may be characterised through different lenses: (i) *ante hoc* interpretable model vs. *post hoc* explanation method, (ii) model or method family, (iii) type of biomedical or healthcare data, (iv) technical goal or problem addressed, and (v) degree of supervision assumed. Chapter 2 has been omitted, as it does not introduce a novel model or method.

Table 1.1 provides a more detailed summary of the scope and structure. In addition to the three aspects discussed above, we report specific model and method classes and the assumed degree of supervision. Motivated by diverse application scenarios, we will study interpretable neural network architectures, prototype- and concept-based models, and concept- and attribution-based explanation techniques. For the uninformed reader, these model and method classes will be explained in Chapter 2.

The scenarios we will investigate assume varying degrees of supervision [35]. In particular, Chapter 3 can be categorised into *self-supervised* learning [36], which leverages intrinsic structure within data for supervision. Chapter 4 is most closely related to the *weakly-supervised* setting [37], where the given labels are partial and noisy. Lastly, Chapters 5–7 will assume *strong supervision* as described by Otálora *et al.* [38], which, in addition to the target variable, requires access to so-called strong labels, e. g., in our case, these are high-level human-understandable attributes, also referred to as *concepts*. While the form of supervision provides a helpful perspective to describe the different learning scenarios, notably, it will not be the primary focus of our discussion throughout this thesis.

In summary, our primary contribution is the **development of interpretable models and explanation techniques with a principled treatment of specific biomedical and healthcare data types to solve application- and user-grounded problems**. Below, we will discuss the contributions of individual chapters. The summaries we provide are relatively nontechnical, and further details can be found in the respective sections.

To begin with, Chapter 2 will provide a scoping review of the recent literature on interpretable and explainable ML. We will define basic concepts and outline a model- and methods-centric perspective on the state of the art. This chapter contributes to the vast survey literature and, in contrast to the reviews with high-level treatment of the topic or limited scope, will focus on *concrete* examples.

In Chapter 3, opening Part I, we will tackle the problem of inferring relations within multivariate time series, also known as *structure learning*, under nonlinear dynamics. Building on the previous literature, we will introduce a class of interpretable neural-network-based models for time series analysis. Due to the nature of the underlying problem, our empirical evaluation will be limited to synthetic data. However, we will propose several plausible applications in biosignal analysis and remote patient monitoring.

Subsequently, Chapter 4 will turn to survival analysis, another problem central to biomedical and healthcare applications. It will introduce a probabilistic generative model to cluster high-dimensional and unstructured survival data. Next to time-to-event prediction, the clustering will help discover outcome- and covariate-driven patient subgroups and devise prototype-based explanations for the nonlinear relationship with the survival time. Alongside simpler benchmarks, we will apply our method to computed tomography images from lung cancer patients.

From Chapter 5 onward, we will concentrate on the more parsimonious concept-based models and methods that leverage high-level attributes instead of directly handling output-input relationships. We will enhance an existing class of concept-based neural network models to scale them up to the typical challenges of medical imaging datasets: the presence of multiple views and the incompleteness of the observed concept variables. The primary application will be the prediction of appendicitis in pediatric patients based on abdominal ultrasound images.

Chapter 6, the first of the two in Part II, will explore concept-based prediction in the *post hoc* setting. We will introduce a simple procedure for interacting with a black-box neural network via concept variables to steer the model's predictions. Moreover, we will investigate various fine-tuning strategies to increase the effectiveness of such concept-based "interventions". In addition to simulated and natural image data, we will evaluate our methods on publicly available chest radiograph datasets.

Finally, Chapter 7 will investigate *post hoc* model editing and the interface between explainable machine learning and algorithmic fairness. Concretely, our goal will be to *debias* black-box models w.r.t. some sensitive attributes, i. e. to mitigate the model's sensitivity to characteristics like ethnicity or gender. We will introduce criteria for pruning neural networks based on the attribution explanation methods and also explore a fine-tuning approach to reduce bias directly. Similar to Chapter 6, we will demonstrate the utility of our methods on chest X-ray classifiers.

1.2 PUBLICATIONS

The contents and text of Chapters 2–7 are based on the following preprints and peer-reviewed publications:[1]

R. Marcinkevičs and J. E. Vogt, *Interpretability and explainability: A machine learning zoo mini-tour*, arXiv:2012.01805, 2020. DOI: 10.48550/arXiv.2012.01805.

R. Marcinkevičs and J. E. Vogt, "Interpretable models for Granger causality using self-explaining neural networks", in *9th International Conference on Learning Representations, ICLR 2021*, OpenReview.net, 2021. DOI: 10.48550/arXiv.2101.07600.

L. Manduchi†, R. Marcinkevičs†, M. C. Massi, T. J. Weikert, A. Sauter, V. Gotta, T. Müller, F. Vasella, M. C. Neidert, M. Pfister, B. Stieltjes, and J. E. Vogt, "A deep variational approach to clustering survival data", in *10th International Conference on Learning Representations, ICLR 2022*, OpenReview.net, 2022. DOI: 10.48550/arXiv.2106.05763.

R. Marcinkevičs, E. Ozkan, and J. E. Vogt, "Debiasing deep chest X-ray classifiers using intra- and post-processing methods", in *Proceedings of the 7th Machine Learning for Healthcare Conference*, Z. Lipton, R. Ranganath, M. Sendak, M. Sjoding, and S. Yeung, Eds., ser. Proceedings of Machine Learning Research, vol. 182, 2022, 504. DOI: 10.48550/arXiv.2208.00781.

R. Marcinkevičs and J. E. Vogt, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples", *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 3, e1493, 2023. DOI: 10.1002/widm.1493.

R. Marcinkevičs†, S. Laguna†, M. Vandenhirtz, and J. E. Vogt, "Beyond concept bottleneck models: How to make black boxes intervenable?", in *NeurIPS 2023 Workshop on XAI in Action: Past, Present, and Future Applications*, 2023. DOI: 10.48550/arXiv.2401.13544.

R. Marcinkevičs†, P. Reis Wolfertstetter†, U. Klimiene†, K. Chin-Cheong, A. Paschke, J. Zerres, M. Denzinger, D. Niederberger, S. Wellmann, E. Ozkan, C. Knorr, and J. E. Vogt, "Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis", *Medical Image Analysis*, vol. 91, 103042, 2024. DOI: 10.1016/j.media.2023.103042.

---

[1] Herein, "†" denotes shared first authorship.

During doctoral studies, the author has also contributed to the articles listed below that are not directly included in the current work:

N. Nowak, T. Gaisl, D. Miladinovic, R. Marcinkevics, M. Osswald, S. Bauer, J. Buhmann, R. Zenobi, P. Sinues, S. A. Brown, and M. Kohler, "Rapid and reversible control of human metabolism by individual sleep states", *Cell Reports*, vol. 37, no. 4, 109903, 2021. doi: 10.1016/j.celrep.2021.109903.

I. Daunhawer, T. M. Sutter, R. Marcinkevičs, and J. E. Vogt, "Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models", in *Pattern Recognition. DAGM GCPR 2020*, Z. Akata, A. Geiger, and T. Sattler, Eds., Cham: Springer International Publishing, 2021, 459. doi: 10.1007/978-3-030-71278-5_33.

R. Marcinkevics†, P. Reis Wolfertstetter†, S. Wellmann, C. Knorr, and J. E. Vogt, "Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis", *Frontiers in Pediatrics*, vol. 9, 2021. doi: 10.3389/fped.2021.662183.

P. Roig Aparicio, R. Marcinkevičs, P. Reis Wolfertstetter, S. Wellmann, C. Knorr, and J. E. Vogt, "Learning medical risk scores for pediatric appendicitis", in *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pasadena, CA, USA: IEEE, 2021. doi: 10.1109/ICMLA52953.2021.00243.

A. H. Hatteland†, R. Marcinkevičs†, R. Marquis, T. Frick, I. Hubbard, J. E. Vogt, T. Brunschwiler, and P. Ryvlin, "Exploring relationships between cerebral and peripheral biosignals with neural networks", in *2021 IEEE International Conference on Digital Health (ICDH)*, 2021, 103. doi: 10.1109/ICDH52753.2021.00022.

J. E. Vogt, E. Ozkan, and R. Marcinkevičs, "Introduction to machine learning for physicians: A survival guide for data deluge", in *Digital Medicine*, Jenny Stanford Publishing, 2023, 3. doi: 10.48550/arXiv.2212.12303.

M. M. Schuurmans, M. Muszynski, X. Li, R. Marcinkevičs, L. Zimmerli, D. Monserrat Lopez, B. Michel, J. Weiss, R. Hage, M. Roeder, J. E. Vogt, and T. Brunschwiler, "Multimodal remote home monitoring of lung transplant recipients during COVID-19 vaccinations: Usability pilot study of the COVIDA desk incorporating wearable devices", *Medicina*, vol. 59, no. 3, 2023. doi: 10.3390/medicina59030617.

M. Vandenhirtz, L. Manduchi, R. Marcinkevičs, and J. E. Vogt, *Signal is harder to learn than bias: Debiasing with focal loss*, arXiv:2305.19671, 2023. doi: 10.48550/arXiv.2305.19671.

R. Marcinkevics†, P. N. Silva†, A.-K. Hankele†, C. Dörnte, S. Kadelka, K. Csik, S. Godbersen, A. Goga, L. Hasenöhrl, P. Hirschi, H. Kabakci, M. P. LaPierre, J. Mayrhofer, A. C. Title, X. Shu, N. Baiioud, S. Bernal, L. Dassisti, M. D. Saenz-de-Juano, M. Schmidhauser, G. Silvestrelli, S. Z. Ulbrich, T. J. Ulbrich, T. Wyss, D. J. Stekhoven, F. S. Al-Quaddoomi, S. Yu, M. Binder, C. Schultheiss, C. Zindel, C. Kolling, J. Goldhahn, B. K. Seighalani, P. Zjablovskaja, F. Hardung, M. Schuster, A. Richter, Y.-J. Huang, G. Lauer, H. Baurmann, J. S. Low, D. Vaqueirinho, S. Jovic, L. Piccoli, S. Ciesek, J. E. Vogt, F. Sallusto, M. Stoffel, and S. E. Ulbrich, "Machine learning analysis of humoral and cellular responses to SARS-CoV-2 infection in young adults", *Frontiers in Immunology*, vol. 14, 2023. doi: 10.3389/fimmu.2023.1158905.

Z. Xiao, M. Muszynski, R. Marcinkevičs, L. Zimmerli, A. D. Ivankay, D. Kohlbrenner, M. Kuhn, Y. Nordmann, U. Muehlner, C. Clarenbach, J. E. Vogt, and T. Brunschwiler, "Breathing new life into COPD assessment: Multisensory home-monitoring for predicting severity", in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23, Paris, France: Association for Computing Machinery, 2023, 84. doi: 10.1145/3577190.3614109.

2

IN PURSUIT OF INTERPRETABILITY AND EXPLAINABILITY

This chapter provides a scoping literature review on interpretable and explainable machine learning, defining key concepts and establishing a taxonomical methods- and model-centric perspective. We will overview a few interpretable model classes and explanation techniques originating from classical computational statistics and delve into more modern yet closely related advances from the machine learning literature. We focus primarily on the supervised learning scenario, as most settings studied in the remainder of the thesis include some form of supervision (Table 1.1), either explicitly or implicitly. This chapter is based on the contents and text of the preprint "Interpretability and Explainability: A Machine Learning Zoo Mini-tour" [39] and publication "Interpretable and Explainable Machine Learning: A Methods-centric Overview with Concrete Examples" [40].

2.1 WHAT'S IN A NAME?

As the name of this chapter suggests, there exist numerous terms often utilised synonymously to describe the desired property of a model or, more broadly, a system, for instance, "*interpretability*" [11], [13], [15], [41], "'*explainability*" [11], [41], [42], "*intelligibility*" [26], [43], or "*understandability*" [13], [44], [45]. Despite the abundance of recent literature broadly attributed to the areas such as interpretable and explainable machine learning, the community has neither agreed on universal definitions for the terms above nor established if there exist any substantial differences among the desiderata that are usually embedded in such terms.

For convenience, within the scope of this thesis, we demarcate interpretable and explainable ML as suggested by Rudin [11]. Namely, interpretable machine learning concerns itself with the models that are interpretable *ante hoc*, i. e. by design, so-called white-, glass-, or grey-box models. In contrast, explainable ML develops techniques and diagnostics that help understand an opaque or black-box model *post hoc*. Beyond this, many criteria could be used to establish taxonomies of interpretable models and

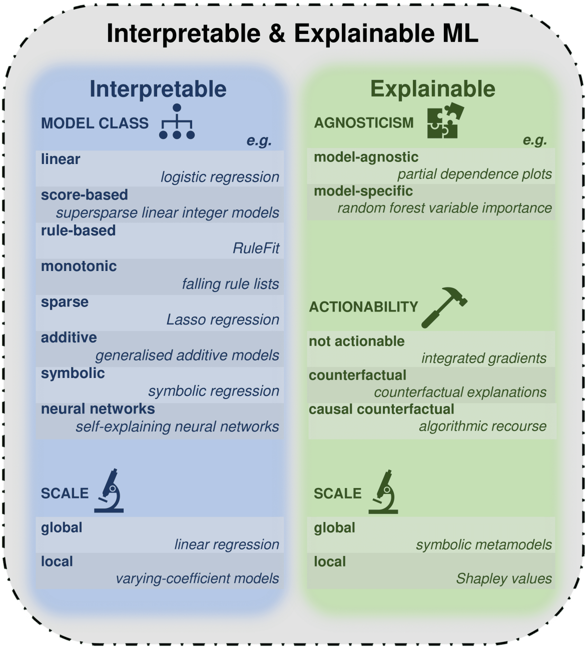


Figure 2.1: Overview of the interpretable and explainable machine learning listing the characteristics commonly used to taxonomise models and methods. Concrete examples are shown in *italics*. Taken from Marcinkevičs and Vogt [40].

explanation techniques [13], [15], [25], [41]. Figure 2.1 lists a few such properties commonly described in the literature.

For example, many interpretable models by design belong to specific classes that possess some properties motivated by the application at hand, for instance, linearity [46], monotonicity [21], [47], or additivity [26], [43], [45], [48]. The scale at which the models are interpreted and explained may range from instance-specific to global. Some works postulate that explanations should be actionable, i. e. they should instruct the user on how to revert algorithmic decisions [12], [49]. Another property of the explanation methods often cited in the literature is agnosticism w.r.t. the model explained [50]: some techniques are tailored to specific model classes, whereas others treat the model as a complete black box. Above, we listed just a few salient properties used to characterise models and methods, and a systematic review is beyond the scope of the current chapter.

In the following sections, we will attempt to pinpoint the problems that interpretable models and explanation techniques address in more technical terms. Furthermore, we will describe specific model and method classes relevant to the following chapters in more detail and with concrete examples.

2.2 PRELIMINARIES AND NOTATION

Firstly, let us introduce the typical supervised learning scenario [35] and some notation used throughout this chapter as a minimal running example. We assume given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where each data point $(\boldsymbol{x}_i, y_i)$ is a tuple of features $\boldsymbol{x}_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$ and, thus, lives in the domain $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$. If the data are tabular, then $\mathcal{X} \subseteq \mathbb{R}^p$ where p denotes the number of features. One of the most well-studied supervised learning problems in the general and interpretable ML literature is classification: in this task, labels are categorical, i. e. $\mathcal{Y} = \{1, \ldots, C\}$ where C is the number of classes. A prevalent special case is $C = 2$, known as binary classification.

In supervised learning, the goal is to find a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ predicting labels from the given features as accurately as possible. The standard approach to learning is to minimise some loss function $\ell : \mathcal{F} \times \mathcal{D} \rightarrow \mathbb{R}^+$ on the given training data, where $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$ and $\ell\left(f, (\boldsymbol{x}, y)\right)$ quantifies the model's prediction error for data point $(\boldsymbol{x}, y)$, e. g. think of the mean squared error (MSE), cross-entropy (CE), or 0-1 loss. The alternative notation we will consider is $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ where the loss for data point $(\boldsymbol{x}, y)$ is given by $\ell\left(f\left(\boldsymbol{x}\right), y\right)$. Herein, the prediction made by the model is denoted by $\hat{y} = f\left(\boldsymbol{x}\right)$. Unless specified otherwise, we will adhere to the setting and notation above. Further notation is introduced on p. xviii.

2.3 INTERPRETABLE MACHINE LEARNING

Rudin [11] argues that a specific all-purpose definition of interpretability cannot exist, as this notion is domain-specific. For example, sparse predictive models are deemed interpretable in genomics data analysis, whereas in medical image analysis, sparsity at the input level, i. e. in the pixel space, is hardly desirable. Despite this, interpretability in the typical supervised learning scenario has been broadly defined through constrained empirical risk minimisation [51], [52].

Definition 2.3.1 [Informal, Interpretable Machine Learning, Dziugaite, Ben–David, and Roy [51] and Rudin *et al.* [52]]**.** Given a domain $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, function class $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$, set of interpretable models $\mathcal{F}_I \subseteq \mathcal{F}$, loss function $\ell : \mathcal{F} \times \mathcal{D} \rightarrow \mathbb{R}^+$, interpretability penalty $\Omega : \mathcal{F} \rightarrow \mathbb{R}^+$, penalty weight $C \in \mathbb{R}^+$, and dataset $\{(x_i, y_i) : 1 \leq i \leq N, (x_i, y_i) \in \mathcal{D}\}$, the interpretable supervised learning setup is given by

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f, (x_i, y_i)\right) + C \cdot \Omega\left(f\right), \text{ subject to } f \in \mathcal{F}_I. \tag{2.1}$$

The definition above augments classical supervised learning by including the interpretability penalty in the objective, representing so-called soft constraints, and introducing hard constraints specified by $\mathcal{F}_I$. Both the penalty and constraints can be adapted according to the domain-specific considerations, and the tradeoff between predictive performance and interpretability can be controlled via parameter C. Although we restrict the discussion to supervised learning, as Rudin *et al.* [52] point out, Definition 2.3.1 is readily applicable to various unsupervised setups where $\mathcal{F}$, $\mathcal{D}$, and ℓ would need to be redefined appropriately.

Beyond reasoning about the interpretability of some model f, there exist works on assessing the simplicity of learning problems [53], trying to understand the tradeoff between predictive performance and interpretability and formulate when interpretable models perform "well enough".

Following the abovementioned setting, interpretable models can be differentiated based on their class, determined by $\mathcal{F}_I$, and penalties that encode soft interpretability constraints. Below, we explore a few model classes and inductive biases common in the interpretable ML literature. These concrete examples are relevant to understanding the general "logic" and prior work behind the new methods introduced in the following chapters of this thesis.

2.3.1 *Rule-based Models and Scoring Systems*

Logic and, specifically, rules are arguably one of the most human-friendly ways of representing our datasets and hypotheses [54], [55]. Rule-based models are usually applied to classification tasks and can be thought of as (ordered) sets of if-then rules that must be evaluated on the given input to arrive at the prediction.

Example 2.3.1 [If-then Rule Lists]**.** Let $\mathcal{X} = \{0, 1\}^2$ and $\mathcal{Y} = \{0, 1\}$. Below is an example of a rule-based classifier consisting of two rules:

R$_1$: IF $x_1 = 0$ AND $x_2 = 0$ THEN $\hat{y} = 1$

R$_2$: IF $x_1 = 1$ AND $x_2 = 1$ THEN $\hat{y} = 1$

Note that a negative prediction ($\hat{y} = 0$) is returned if no rules apply for the given feature vector $\boldsymbol{x}$.

While a single or a few if-then rules are easy to understand and evaluate by hand, unregularised rule lists can quickly become incomprehensible in datasets with many features and complex relationships. Therefore, some works have focused on learning better structured, i. e. more interpretable, lists. For example, decision trees [56] can be seen as an instance of rule lists admitting a simple visual representation. Some research efforts have focused on explicitly reducing the number of rules [57], i. e. pruning. Another line of work has developed a method to convert sparse linear models with pairwise feature interactions into rule lists [58]. Further improvement includes learning lists where the rules are ordered monotonically according to the predicted probability of the positive label [21], [59], [60].

Scoring systems are another class of parsimonious models related to rule-based approaches. In the healthcare domain, risk scores are ubiquitous, e. g. the APACHE score [61] used for classifying the disease severity or the SOFA score [62] for assessing organ failure in intensive care unit (ICU) patients. Most of these have been designed *ad hoc* based on the domain expert knowledge, i. e. not in a data-driven manner. In contrast, supersparse linear integer models (SLIM) [19], [20] are linear classification models with integer-valued parameters which mimic the format of the conventional risk scores yet are learnt from the data (Definition 2.3.2).

Definition 2.3.2 [Supersparse Linear Integer Models, Ustun and Rudin [19]]**.** Consider $\mathcal{X} = \{0, 1\}^p$ and $\mathcal{Y} = \{-1, 1\}$. Let $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x}$ be a supersparse linear integer model. Its optimal parameters are defined by the following integer linear programme (ILP):

$$\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i \leq 0\}} + \lambda_0 \left\|\boldsymbol{\beta}\right\|_0 + \lambda_1 \left\|\boldsymbol{\beta}\right\|_1, \text{ subject to } \boldsymbol{\beta} \in \mathcal{B}, \tag{2.2}$$

where $\mathbf{1}$ denotes an indicator function, $\mathcal{B} = \{-L, -L + 1, \ldots, U - 1, U\}^p$, $L, U \in \mathbb{N}$, and $\lambda_0, \lambda_1 > 0$ are penalty term weights.

Observe that per Definition 2.3.2, SLIMs have integer-valued coefficients $\boldsymbol{\beta} \in \mathcal{B}$, and the predicted score given by $\boldsymbol{\beta}^\top \boldsymbol{x}$ is a sum of integral coefficients for the nonzero features in $\boldsymbol{x}$. Also note that the objective in Equation 2.2 includes penalties on both the ℓ_0- and ℓ_1-norms of the coefficient vector, and, hence, the optimal parameters will be sparse and the model itself consequently more interpretable (Section 2.3.2). Another essential advantage of the ILP formulation above is that domain-specific considerations can be readily incorporated by specifying additional constraints [19]. Recent work has focused on learning even more parsimonious predictive models by solving similar ILPs, for example, to construct predictive checklists [63].

In summary, while if-then rules offer a simple representation of the data and possible hypotheses, many additional inductive biases can be incorporated into rule-based models to improve their interpretability, e. g. sparsity, linearity, and monotonicity. Other related model classes are risk scores and predictive checklists, whose format is loosely inspired by medical scoring systems. The apparent limitation of such models is their reliance on discretised tabular features and, hence, the necessity for careful feature engineering when working with unstructured data types.

2.3.2 *Sparsity-inducing Regularisation*

As mentioned above regarding SLIMs, sparsity helps improve the interpretability of the model and data representation [64], [65]. When referring to supervised learning and linear models in particular, sparsity usually implies having few nonzero parameters, i. e. the model relies on fewer features. Such a constraint can be imposed by introducing an ℓ_0-norm penalty into the optimisation objective (Equation 2.2). Optimisation under such regularisation can be computationally infeasible, so various approximations have been introduced.

A standard approach is to replace the ℓ_0-norm with its convex relaxation. For instance, Lasso (least absolute shrinkage and selection operator) constrains the ℓ_1-norm of the coefficient vector [66]. This classical regularisation technique has seen many extensions and practical applications in generalised linear models (Definition 2.3.5), e. g. in logistic [67] and Cox [68] regression.

Definition 2.3.3 [Lasso Regression, Tibshirani [66]]. Consider $\mathcal{X} = \mathbb{R}^p$. Let $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x}$ be a linear model. Then, the Lasso estimator is given by solving the following problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f\left(\boldsymbol{x}_i\right), y_i\right) + \lambda \left\|\boldsymbol{\beta}\right\|_1, \tag{2.3}$$

where $\lambda > 0$ is the penalty term weight.

Below, we will briefly introduce a couple of generalisations of the objective and the problem setting from Definition 2.3.3. To address high-dimensional regression problems, i. e. cases where $p \gg N$, the elastic net regularisation [69] combines Lasso and ridge penalties into $\Omega\left(f\right) = \left(1-\alpha\right) \left\|\boldsymbol{\beta}\right\|_1 + \alpha \left\|\boldsymbol{\beta}\right\|_2^2$, where $\alpha \in (0, 1)$ controls the relative weights of the Lasso and ridge terms.

Another noteworthy setting explored in the literature is the case of grouped features [70], [71], wherein the features are arranged into nonoverlapping groups. The group Lasso regularisation (Definition 2.3.4) yields the coefficients that are sparse w.r.t. the groups, i. e. the weights are all either zero or nonzero for all features within a specific group.

Definition 2.3.4 [Group Lasso Regression, Yuan and Lin [71]]. Consider $\mathcal{X} = \mathbb{R}^p$. Assume given M feature groups $\mathcal{G}_1, \ldots, \mathcal{G}_M$ s.t. $\bigcup_{j=1}^{M} \mathcal{G}_j = \{1, \ldots, p\}$ and $\mathcal{G}_j \cap \mathcal{G}_k = \emptyset$ for all $1 \leq j \neq k \leq M$. The group Lasso estimator is obtained by solving

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f\left(\boldsymbol{x}_i\right), y_i\right) + \lambda \sum_{j=1}^{M} \left\|\boldsymbol{\beta}_{\mathcal{G}_j}\right\|_2, \tag{2.4}$$

where $\lambda > 0$ is the penalty term weight.

In addition to the overall penalty weight λ, group-specific coefficient norms $\left\|\boldsymbol{\beta}_{\mathcal{G}_j}\right\|_2$ may be weighted individually, for instance, according to the group sizes or domain knowledge and considerations. Moreover, the approach above is readily extensible to groups that may overlap, i. e. $\mathcal{G}_j \cap \mathcal{G}_k \neq \emptyset$ for some (j, k). Generally, scenarios beyond the standard ℓ_1-norm regularisation (Equation 2.3) are known as structured sparsity [72] and assume that features are "organised" into some form of groups (Equation 2.4), networks, or other structures. Adaptive methods even exist that facilitate learning both the parameters and group structure in bilevel optimisation [73].

Overall, sparsity is an inductive bias instrumental in learning parsimonious and hence, interpretable predictive models. Sparsity can be easily formalised and attained for classes other than rule lists or risk scores (Section 2.3.1), for example, as outlined above, for generalised linear models. In the later chapters, we will demonstrate that sparsity can also be helpful in more complex model classes. Nevertheless, for sparsity to be meaningful, it must be defined w.r.t. some human-understandable and preferably high-level features, and the model must be able to disentangle the parameters and "contributions" belonging to the individual inputs [74].

2.3.3 *Generalised Linear, Generalised Additive, and Varying-Coefficient Models*

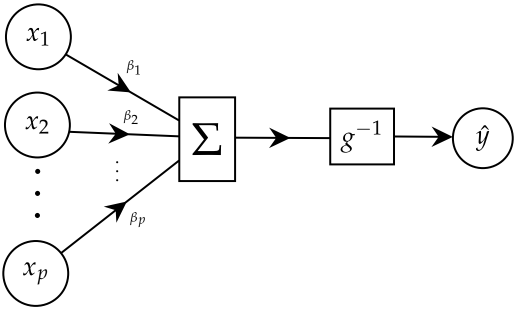
Many conventional modelling techniques from statistics can be deemed interpretable, as they come with a battery of diagnostics and facilitate direct interpretation of their parameters, e. g. the effect of an individual feature in linear regression can be read off directly from its coefficient [75]. Generalised linear models (GLM) [76], [77] form a bedrock of modern regression modelling. As evidenced by the name, GLMs are a general class with linear, logistic, and Poisson regression as special cases.

Definition 2.3.5 [Generalised Linear Models, Nelder and Wedderburn [76]]**.** Consider $\mathcal{X} = \mathbb{R}^p$ and let g be an appropriately chosen monotonically increasing function. A generalised linear model is given by

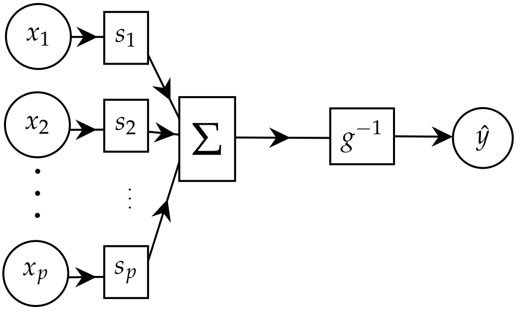
$$f\left(\boldsymbol{x}\right) = g^{-1}\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right), \tag{2.5}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are regression coefficients. $\boldsymbol{\beta}^\top \boldsymbol{x}$ is usually called the linear predictor, and g is the link function.

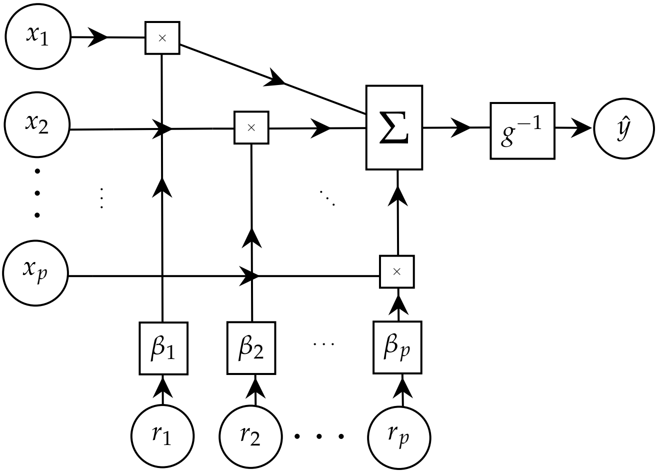
Below, we will consider two extensions of GLMs that can capture more sophisticated relationships than those given by Equation 2.5. Nevertheless, these extended classes still adhere to the interpretable format from Definition 2.3.5, in that contributions and parameters belonging to individual features can be disentangled. This characteristic of the GLMs and their extensions is well apparent from the schematic summaries shown in Figure 2.2.



(a) GLM



(b) GAM



(c) VCM

FIGURE 2.2: Schematic representation of the (a) generalised linear models (GLM) and their extensions: (b) generalised additive models (GAM) and (c) varying-coefficient models (VCM). Circular and rectangular nodes correspond to variables and functions, respectively.

One reasonable modification is to allow nonlinearities in individual variables while still keeping the input to g^{-1} additive—generalised additive models (GAM) [48] build on this idea of representing the response variable as a nonlinear but additively separable function [78] of features (Figure 2.2b).

Definition 2.3.6 [Generalised Additive Models, Hastie and Tibshirani [48]]**.** A generalised additive model is given by

$$f\left(\boldsymbol{x}\right)=g^{-1}\left(\sum_{j=1}^{p}s_{j}\left(x_{j}\right)\right), \tag{2.6}$$

where s_j are smooth learnable functions, sometimes called shape functions [43].

Typically, s_j are estimated using smoothing splines [79] whose smoothness is controlled by including the penalty term $\Omega\left(f\right)=\sum_{j=1}^{p}\int\left[s_{j}^{\prime\prime}\left(x\right)\right]^{2}dx$ into the loss function. A recent line of works [26], [43], [45] argues in favour of using non-smooth shape functions, e. g. boosted stumps and trees, especially to capture relationships with abrupt changes. Beyond univariate nonlinearities, Lou *et al.* [45] explore GAMs with pairwise interactions among the features by introducing terms $s_{j,k}\left(x_j,x_k\right)$ for $1\leq j\neq k\leq p$ in Equation 2.6, thus, trading off complete additive separability for increased model flexibility. In addition to the smoothness of shape functions, other inductive biases can be incorporated into GAMs. For instance, Liu *et al.* [80] propose sparse additive models by combining GAMs with sparse linear modelling (Definition 2.3.3).

Like GAMs, varying-coefficient models (VCM) [81]–[83] are a generalisation of GLMs (Definition 2.3.5). Instead of smooth nonlinearity in features, VCMs allow for smoothly-varying coefficients. In addition to the feature vector $\boldsymbol{x}$, these models depend on another set of variables $\boldsymbol{r}$, referred to as "*effect modifiers*", that modulate the coefficients of the individual features (Figure 2.2c).

Definition 2.3.7 [Varying-coefficient Models, Hastie and Tibshirani [81]]**.** Varying-coefficient models have the following form:

$$f(\boldsymbol{x},\boldsymbol{r})=g^{-1}\left(\sum_{j=1}^{p}x_{j}\beta_{j}(r_{j})\right), \tag{2.7}$$

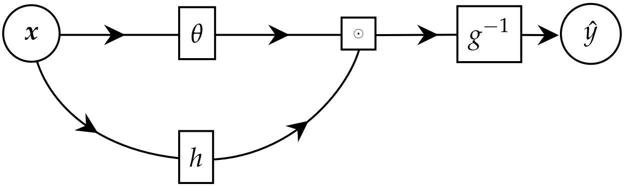
where β_j are smooth coefficient functions mapping effect modifiers r_j to varying coefficients for $1\leq j\leq p$.

In practice, the effect modifiers $\boldsymbol{r}$ can be represented by the covariates $\boldsymbol{x}$ themselves or by some exogenous factors, such as time or space. The formulation in Definition 2.3.7 arises naturally from many real-world modelling problems, e. g. in longitudinal and spatiotemporal data. VCMs are an extension of the two classes presented above (Equations 2.5–2.6). On the one hand, when the effect modifiers are constant, Equation 2.7 is equivalent to the definition of the GLM (Equation 2.5); on the other hand, when the *features* are constant, VCMs become equivalent to GAMs (Equation 2.6). Observe that, generally, without further assumptions, as opposed to GAMs, VCMs are not additively separable and, hence, can model high-order feature interactions.

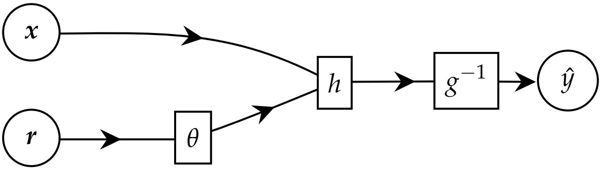
To summarise, generalised linear models and their extensions allow modelling the relationship between features and labels via a monotonic link function and provide a unifying perspective on several classical regression models. GLMs still follow an interpretable format wherein the inverse of the link function is applied to the linear predictor. GAMs and VCMs build on the interpretable form of GLMs and incorporate nonlinearities, relying on additivity and smoothness constraints to keep a certain degree of local interpretability. Although similar to linear and sparse models (Section 2.3.2), GLMs and their extensions rely on having a comprehensible feature space; several recent works have explored related model classes that could be applied to unstructured data [84]–[87].

2.3.4 *Interpretable Neural Networks*

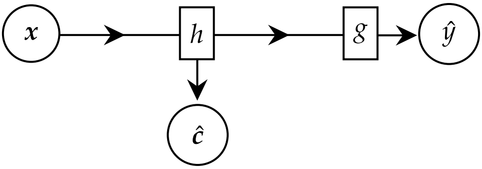
The last three decades of machine learning research and practice have seen immense progress brought about by advancements in deep learning [6], [7]. The prolific growth in model complexity and size [88] sparks an even greater interest in and creates a more urgent demand for developing interpretable neural network (NN) models. Below, we will overview a few recent lines of research that, in some manner, tackle this problem. Figure 2.3 shows schematic sketches of several salient network architectures from the works mentioned in the current section.



(a) SENN



(b) CEN



(c) CBM

FIGURE 2.3: Simplified schematic representation of the (a) self-explaining neural networks (SENN), (b) contextual explanation networks (CEN) and (c) concept bottleneck models (CBM). Circular and rectangular nodes correspond to variables and functions, respectively; bold letters denote vector variables.

In Sections 2.3.1–2.3.2, we have discussed sparsity-inducing regularisation in the context of linear models. Similar techniques have been explored in application to various neural network architectures. For instance, Feng and Simon [89], [90] thoroughly study, theoretically and empirically, sparse-input neural networks (SPINN) and ensembles composed thereof. The backbone of SPINNs is a standard fully connected neural network, or multilayer perceptron (MLP). However, the loss function of SPINNs includes the sparse-group Lasso penalty [91], an extension of the group Lasso from Definition 2.3.4, on the input-layer weights grouped by the input variable. Thus, thanks to the sparsity-inducing regularisation, the first layer of a SPINN effectively performs feature selection. Several concurrent works investigate similar approaches, such as deep feature selection [92], which applies a sparse mask at the input of a neural network, and LassoNet [93], which, in addition to the input-layer sparsity, introduces a linear skip layer and utilises a modified hierarchical penalty motivated by the hierarchy principle [94]. Beyond the supervised learning scenario, similar model design principles have been leveraged in time series analysis, e. g. Tank *et al.* [95] and Khanna and Tan [96] focus on recurrent neural network architectures, adapting SPINNs to autoregressive time series modelling.

Another research direction can be loosely described as developing neural networks mimicking the structure of the generalised linear models and their generalisations (Section 2.3.3). These extensions usually amortise some of the model parameters using neural nets; in the context of Figure 2.2, this amounts to replacing some of the nodes in the computational graph with NNs. Two representative examples are the models proposed by Alvarez Melis and Jaakkola [84] and Al-Shedivat, Dubey, and Xing [85], both of which can be seen as neural-network-based extensions of the VCMs (Definition 2.3.7). In particular, self-explaining neural networks (SENN) by Alvarez Melis and Jaakkola [84] combine two neural networks outputting varying coefficients and mapping raw features to high-level concepts: $f(\boldsymbol{x}) = g^{-1}\left(\theta\left(\boldsymbol{x}\right) \odot h\left(\boldsymbol{x}\right)\right)$ (Figure 2.3a). Contextual explanation networks (CEN; Figure 2.3b) by Al-Shedivat, Dubey, and Xing [85] cast VCMs into a deep probabilistic setting where the label is assumed to be generated conditionally on the features and parameters, in their turn, conditioned on some context variables, i. e. effect modifiers ($\boldsymbol{r}$ in Equation 2.7). Similarly, another line of work combines GAMs and neural networks [86], [87]. The motivation behind these works is the improvement in the interpretability of neural networks and the scalability of the GAMs, which in their original form are not fully differentiable and, hence, cannot be readily incorporated into modular deep learning frameworks.

Beyond the classical statistical perspective on the interpretability of NNs, many works explore the utility of the attention mechanism [97]–[99], which can help understand relationships between the network's inputs and outputs and has become a workhorse of many modern models for natural language processing and sequential data. For example, Schwab, Miladinovic, and Karlen [100] introduce the attentive mixture of experts models that equip a mixture of connected expert subnetworks with a regularised attention mechanism, which can help evaluate feature importance. Nauta, Bucur, and Seifert [101] leverage attention to a similar end to infer time series structure using convolutional neural networks. Despite its popularity and practicality, some works have scrutinised the use of the unregularised attention mechanism for interpreting NNs [102], [103], observing that attention weights often disagree with other feature importance measures, such as output-input gradients (Section 2.4).

Most model designs mentioned so far focus mainly on predictive modelling and seek interpretability in the raw feature space. A different approach, prevalent in the literature, concentrates on representation learning [104]. Although many of the techniques below do not explicitly postulate interpretability as a desideratum, arguably, it is among the ultimate goals of representation learning. For instance, many generative models try to learn disentangled representations in an unsupervised and weakly- or semi-supervised manner [105]–[110]. Ideally, the components of such representations should be uniquely correlated with the ground-truth factors of variability. Apart from generative modelling, there is a renewed interest in concept-based prediction models [22], [111]–[114], which split the prediction into two consecutive steps (Figure 2.3c): (i) high-level human-understandable concept prediction based on the input features and (ii) label prediction based on the concept values predicted in the first step, i. e. a classifier f can be decomposed as $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$, where h maps input features to the concepts. An expert user can understand predictions made by such models by examining the concepts and may even interact with the model by editing its intermediate representations. Thus, in addition to mere model "introspection", interpretations can help bring humans into the algorithmic decision-making loop.

This section has given a general scoping overview of the numerous attempts at designing interpretable neural networks with a specific focus on the trends that are related to and have inspired the work described in the following chapters of this thesis. Although the methodological advancements are seemingly abundant, many of them do not venture beyond the classification task and disregard more specialised problem settings, e. g. clustering, multiview and multimodal learning, time series and survival analysis, or interactive learning. Thus, many technical problems at the interface of interpretable ML and well-established application areas remain unaddressed or under-explored, especially for deep learning.

2.4 EXPLAINABLE MACHINE LEARNING

Although *ante hoc* interpretability is conceptually compelling, there are many practical reasons why interpretability "by design" may not always be feasible. For instance, the model may be proprietary, and its parameters may not be made accessible to the users; or, at the development time, the concerns about interpretability did not arise and, hence, were not addressed, whereas redesigning and retraining the model might be costly or even impossible. Explainable ML, sometimes also broadly referred to as explainable artificial intelligence (XAI) [115], [116], studies methods and algorithms for explaining black-box models *post hoc* [11], i. e. after they have been trained. Explanations can take on various forms, e. g. visual, symbolic, or textual, and, in some instances, many of them could be thought of as model diagnostics [117], summarising which parts of the given input or which data points from the training set have contributed to the prediction.

This section will briefly summarise several well-established explanation method families, specifically, the techniques developed for or applicable to NN models. Needless to say, many methods exist beyond the scope of deep learning, originating from classical statistical learning research, e. g. partial dependence plots [5] and feature importance measures [4].

2.4.1 *Attribution Methods*

Attribution methods offer a practical approach to explaining output-input relationships and are used ubiquitously as a model diagnostic for deep neural networks in applied research. For instance, works by Kelley *et al.* [118] and Arcadu *et al.* [119] are typical examples from the biomedical literature.

Definition 2.4.1 [Informal, Attribution, Sundararajan, Taly, and Yan [120]]**.** Consider $\mathcal{X} = \mathbb{R}^p$. Given a black-box model $f : \mathbb{R}^p \rightarrow \mathbb{R}$, an input $\boldsymbol{x} \in \mathcal{X}$, and a reference input $\boldsymbol{x}_0 \in \mathcal{X}$, an attribution for the prediction $\hat{y} = f(\boldsymbol{x})$ at input $\boldsymbol{x}$ relative to the baseline $\boldsymbol{x}_0$ is a vector $A_f(\boldsymbol{x}, \boldsymbol{x}_0) = \begin{pmatrix} a_1 & \cdots & a_p \end{pmatrix}^\top$ where a_j quantifies the "contribution" of x_j to $\hat{y}$.

Definition 2.4.1 provides an informal characterisation of attributions, and the technical details on how specifically the function A_f is defined and what properties it satisfies vary from one attribution method to another. Also, note that not all attribution methods require a baseline input; therefore, A_f

may be constant w.r.t. x_0. Typically, x_0 is a signal-free input, such as the zero vector or the empirical average computed across the training set.

To illustrate this informal definition using a concrete example, let us consider an attribution measure that has inspired many follow-up works in modern literature. Shapley value regression [121], [122] leverages a game-theoretic approach to attribution based on Shapley values [123]. Recent works [124] have introduced model-agnostic approximations of Shapley value analysis that are computationally scalable and can be applied to neural network models. Similarly, there exist model-specific methods for computing Shapley values, e. g. in tree-based models [125].

Definition 2.4.2 [Shapley Value Regression, Lundberg and Lee [124]]**.** Given an input $x \in \mathbb{R}^p$, Shapley value for the j-th feature x_j is given by

$$\sum_{\mathcal{S}\subseteq\{1,\ldots,p\}\setminus\{j\}} \frac{|\mathcal{S}|!\,(p-|\mathcal{S}|-1)!}{p!}\left[f_{\mathcal{S}\cup\{j\}}\left(x_{\mathcal{S}\cup\{j\}}\right)-f_{\mathcal{S}}\left(x_{\mathcal{S}}\right)\right], \tag{2.8}$$

where $f_{\mathcal{S}} \in \mathcal{F}$ is the model trained on the inputs with features from $\mathcal{S}$.

As seen from Equation 2.8, the Shapley value for a given regression model (f) and input (x) quantifies the expected change in the predicted value after removing the feature of interest (j) from among the inputs. Although a sensible measure of the feature's contribution, the naïve definition in Equation 2.8 requires retraining the model on exponentially many subsets of the feature set $\{1,\ldots,p\}$ and, therefore, in practice, requires approximation.

Beyond the sensitivity analysis performed by removing the covariate of interest and retraining the model, many attribution techniques have been developed for differentiable models, including NNs, and, thus, utilise output-input gradient information to measure feature contributions. The most straightforward gradient-based statistic is the partial derivative of the function f w.r.t. the feature of interest evaluated at the given data point x' [126]: $\left.\frac{\partial f(x)}{\partial x_j}\right|_{x'}$, which is also known as the gradient-only saliency. Many subsequent methods have addressed the limitations of the vanilla gradient, e. g. making attributions sharper by multiplying gradients with the inputs [127] or building theoretical connections with the cooperative game theory [120]. The latter work, for example, proposes integrating the gradient along the line between the given input data point and reference (Definition 2.4.3). It can be shown that such a measure is closely related to the Aumann–Shapley values proposed in the context of infinite games [128].

Definition 2.4.3 [Integrated Gradients, Sundararajan, Taly, and Yan [120]]**.** For the data point $\boldsymbol{x}$, reference $\boldsymbol{x}_0$, and j-th feature, the integrated gradient is given by

$$\left(x_j - x_{0,j}\right) \int_{\alpha=0}^{1} \frac{\partial f\left(\boldsymbol{x}_0 + \alpha\left(\boldsymbol{x} - \boldsymbol{x}_0\right)\right)}{\partial x_j} d\alpha. \tag{2.9}$$

Similar to the attention (Section 2.3.4), rather than mechanistically explaining *how* the model works, attributions highlight *what* the model focuses on. As for the attention, plenty of counterarguments against the use of attribution have been voiced, e. g. the unfaithfulness to the original model's behaviour [11], the reliance on the raw feature space and, thus, the lack of high-level and conceptual insights [129], the lack of contrast [130], and limited usefulness at detecting unknown spurious correlation [131]. Other explanation method families discussed in the following sections partially address some of the mentioned limitations.

2.4.2 *Surrogate Models*

An alternative to deriving explanations directly from the given black-box model f is to train one or several interpretable models from some prespecified model class $\mathcal{F}_I$ (Definition 2.3.1) to mimic the black box faithfully and explain its predictions. This trick is known as meta- or surrogate modelling [41]. As suggested above, the black box could be either approximated globally, i. e. for all $\boldsymbol{x} \in \mathcal{X}$, using a single interpretable surrogate or locally with many surrogate functions, i. e. separately for small subsets of the instance space. Below we provide concrete examples to illustrate both approaches.

In global surrogate modelling, a single surrogate function f' is optimised to approximate f. Hence, the problem can be formalised as

$$\min_{f' \in \mathcal{F}_I} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f'\left(\boldsymbol{x}_i\right), f\left(\boldsymbol{x}_i\right)\right). \tag{2.10}$$

Thus, instead of training an interpretable model on the ground-truth labels, we train f' to predict the output of f. Since f' belongs to a class of interpretable models, its predictions are explainable "by definition." Herein, like in interpretable machine learning, the primary design choice lies in the specification of $\mathcal{F}_I$. For instance, one viable option is symbolic metamodels

[132] given by a set of succinct mathematical expressions. From a technical perspective, symbolic metamodelling amounts to symbolic regression [133], for which various practical methods have been described in the literature. For instance, Alaa and van der Schaar [132] introduce an elegant differentiable approach based on Meijer G-functions.

While explaining a black box by a single interpretable model is compelling, $\mathcal{F}_I$ might be very different from the original model class and, thus, a single $f' \in \mathcal{F}_I$ might be a poor approximation of the original model. A more flexible approach is to approximate the given black box locally, i. e. using many surrogate functions, as proposed by Ribeiro, Singh, and Guestrin [50] under the term of local interpretable model-agnostic explanations (LIME). In LIME, a surrogate is trained for every given input $\boldsymbol{x}$ to approximate the black box in its neighbourhood.

Definition 2.4.4 [Local Interpretable Model-agnostic Explanations, Ribeiro, Singh, and Guestrin [50]]**.** Given a back-box model f, input $\boldsymbol{x}$, and a class of interpretable models $\mathcal{F}_I$, a local model-agnostic explanation is given by

$$\arg\min_{f' \in \mathcal{F}_I} \ell\left(f, f', \pi_{\boldsymbol{x}}\right) + \Omega\left(f'\right), \tag{2.11}$$

where $\ell\left(f', f, \pi_{\boldsymbol{x}}\right)$ is the loss for approximating f with f' within the neighbourhood of $\boldsymbol{x}$ defined by the proximity measure $\pi_{\boldsymbol{x}}$. Note the difference in the notation compared to the previous uses of ℓ. Similar to Definition 2.3.1, $\Omega : \mathcal{F}_I \rightarrow \mathbb{R}^+$ is a complexity penalty.

Ideally, $\mathcal{F}_I$ from Equation 2.11 should be selected from simpler and computationally scalable classes, e. g. linear Lasso regression models (Definition 2.3.3) or GAMs (Definition 2.3.6) so that the per-data-point optimisation problem above becomes easier. Similar to the global surrogates, an explanation can be derived for each data point from the local surrogate function f'. For example, in the case of the Lasso regression, an explanation is given by sparse regression coefficients (Equation 2.3).

In summary, local and global surrogate modelling resorts to learning another interpretable model to explain the predictions made by a black box. In contrast to gradient-based attribution, surrogate explanations are tied to a concrete interpretable model used for approximation. Nevertheless, metamodelling is limited in so far as an interpretable model of choice can approximate the original model: if the approximation is not faithful, the explanations may be inaccurate. A fair criticism is that one should develop an interpretable model from the beginning instead of approximating a black box with a glass box. Nevertheless, interpretable ML is not always viable, especially when model training requires special hardware, e. g. a graphics processing unit, unavailable at deployment time or when the training data and the original model are proprietary and can only be queried without accessing parameters or training data points.

2.4.3 *Concept-based Explanations*

Most attribution techniques and surrogate models attempt to explain output-input relationships and, thus, mainly define some form of feature importance in the input space. Similar to interpretable representation learning and concept-based models (Section 2.3.4), an alternative to explaining output-input relationships between $\hat{y}$ and $\boldsymbol{x}$ is to use high-level attributes, or concepts, usually represented by an additional variable $\boldsymbol{c} \in \mathcal{C}$. Most prior works focus on categorical and binary concepts, i. e. $\mathcal{C} = \{0, 1\}^K$, where K is the number of concept variables.

Kim *et al.* [129] introduce a statistic quantifying the sensitivity of neural network's representations to a given concept variable and introduce a procedure to globally test whether an NN model utilises the given concept in its predictions. To briefly introduce the idea behind their approach, let us assume observing a single concept variable $c \in \{0, 1\}$ in addition to the covariates $\boldsymbol{x}$ and label y. Below is the definition of conceptual sensitivity, which quantifies how sensitive a neural network's prediction is to the given attribute at a specific layer.

Definition 2.4.5 [Conceptual Sensitivity, Kim *et al.* [129]]**.** Given a neural network f and its slice $\langle g, h \rangle$ where $f = g \circ h$ and h outputs the activation in the layer defined by the slice, for the concept variable $c \in \{0, 1\}$ and an input $\boldsymbol{x} \in \mathcal{X}$, the conceptual sensitivity of f is given by

$$\frac{\partial g\left(h\left(\boldsymbol{x}\right)\right)}{\partial \boldsymbol{v}_c} = \nabla g\left(h\left(\boldsymbol{x}\right)\right)^\top \boldsymbol{v}_c, \tag{2.12}$$

where $\boldsymbol{v}_c$ is the normal to a hyperplane separating data points with $c = 0$ and 1, referred to as the concept activation vector (CAV).

Note that the conceptual sensitivity, as defined in Equation 2.12, corresponds to the directional derivative of the function g w.r.t. the CAV $\boldsymbol{v}_c$, i. e., intuitively, it measures the rate of change in g in the direction specified by the normal to the concept-separating hyperplane in the output space of

h. This statistic can be utilised to assess the global influence of the given concept on the output of the network by, for instance, inspecting the sign of the directional derivative across all data points from one of the classes.

As for the concept-based predictive models, an apparent limitation of conceptual sensitivity is its reliance on the *a priori* known concept variables and the need for annotated instances to train concept classifiers and extract CAVs. Therefore, recent research efforts have been directed at automated concept discovery as a preliminary step to model explanation [134], [135]. Another limitation is the use of the directional derivative and, hence, the assumption that function g is differentiable. Thus, similar to gradient-based attributions, e. g. integrated gradients (Definition 2.4.3), conceptual sensitivity is not a wholly model-agnostic explanation technique.

2.4.4 *Prototype- and Case-based Explanations*

So far, all explanation techniques discussed in this chapter compute some form of importance value for either raw features or high-level attributes. By contrast, prototype- and case-based explanation methods usually output a data point, a set thereof, or summary statistics of several data points to explain the prediction for a specific example. These approaches are best described as instances of case-based reasoning, which "*uses old experiences to understand and solve new problems*" [136].

One of the simplest examples of case-based explanation is the nearest neighbour algorithm, whose prediction for an instance can be explained by the training data point, which is the nearest neighbour. Then, a simple strategy to explain the output of a black-box model f for an input $\boldsymbol{x}$ is to return the nearest neighbour of $\boldsymbol{x}$ given by $\arg\min_{\boldsymbol{x}'} d\left(\boldsymbol{x}, \boldsymbol{x}'\right)$, where d is a distance metric defined by f [137]. For example, if f is a neural network, we could use the Euclidean distance applied to the activation vectors from the network's intermediate layer. In the same vein, for the classification task, the prediction may be additionally explained by the nearest miss, i. e. the nearest training data point belonging to the class different from the predicted one.

A more sophisticated approach is exemplified by the Bayesian case model (BCM) proposed by Kim, Rudin, and Shah [138]. The BCM leverages a generative mixture model to perform probabilistic unsupervised clustering on the training set. Each cluster discovered is characterised by a *prototype*, a training instance with the maximum probability conditional on the cluster's parameters. These prototypes can be used to explain all instances assigned to a specific cluster. In addition, the BCM selects feature subsets that are relevant to each cluster.

Beyond prototype and feature selection, related works have explored the Bayesian model criticism (BMC) framework [139]. In particular, Kim, Khanna, and Koyejo [139] leverage maximum mean discrepancy (MMD) [140] to, in addition to the prototypes, select *criticisms*—data points, which are poorly explained by the chosen prototypes, belonging to the regions of the data space that "deviate" from the prototypes the most.

Case- and prototype-based explanations can compellingly help explore the data space and decision boundary, illustrating those with concrete examples, e. g. nearest neighbours and centroids. Nevertheless, case-based explanations possess some inherent limitations [141]. For instance, they can be sensitive to the choice of the dissimilarity measure. Moreover, such explanations are restricted because they utilise data points to explain the model's predictions. As for the attribution measures (Section 2.4.1), if the raw data space is very high-dimensional and incomprehensible, then examples, prototypes, and criticisms will be of limited use to a human subject.

2.4.5 *Counterfactual Explanations*

Counterfactual explanations [12] form another class of methods, which, similar to case-based explanations (Section 2.4.4), help explore the data space and learnt decision boundary. Intuitively, counterfactual explanations address the question [40]: "*Why this specific prediction was made instead of another?*" Thus, they are meant to be contrastive, actionable, and more human-friendly [25], possibly even offering algorithmic recourse [49].

Below, we provide a formal definition of counterfactual explanations slightly adapted from the work by Wachter, Mittelstadt, and Russell [12].

Definition 2.4.6 [Counterfactual Explanations, Wachter, Mittelstadt, and Russell [12]]**.** Given a back-box model f, an input $x \in \mathcal{X}$, and an alternative target value $y' \in \mathcal{Y}$, a counterfactual explanation is given by

$$\arg\min_{x' \in \mathcal{X}} \; d\left(x, x'\right) + \lambda \ell\left(f\left(x'\right), y'\right), \tag{2.13}$$

where d is the distance function on the feature space and $\lambda > 0$ controls the tradeoff between the proximity and validity of the explanation [142].

The definition above corresponds to the search for a data point that is similar to the given $\boldsymbol{x}$ (*proximity*) yet has a different predicted outcome given by $y' \in \mathcal{Y}$. Herein, $\lambda > 0$ defines the slackness on the constraint $f\left(\boldsymbol{x}'\right) = y'$ (*validity*). Interestingly, counterfactual explanations bear some similarity to the nearest miss from the nearest neighbour algorithm (Section 2.4.4) and adversarial perturbations [143].

Beyond Definition 2.4.6, researchers have explored many variations and extensions of counterfactuals, for instance, producing multiple and diverse explanations [142] and utilising generative models to find higher-fidelity counterfactuals [144]–[146].

While counterfactuals provide contrastive explanations and, in addition to exploring the feature space, do not disregard the decision boundary of the black-box model being explained, their limitations are similar to those of the case-based techniques. Moreover, it is debatable whether, in their original version, counterfactuals are truly actionable because their definition lacks causal perspective, for which the user should resort to algorithmic recourse [49].

2.5 ON APPLICATIONS TO HEALTHCARE

In the previous sections, we have provided an overview of selected models and methods from the "zoo" of interpretable and explainable machine learning. Natural follow-up questions arising from this discussion are (i) how such techniques are being applied to the healthcare data in the current research practices and (ii) whether there are any methodological challenges uniquely associated with the healthcare domain, i. e. what makes interpretable and explainable ML for healthcare "special"? Below, we will briefly delve into these questions, which are relevant to the general topic of this thesis.

While applications of interpretability and explainability can be found across all subfields of healthcare, for concrete references, see the survey by Stiglic *et al.* [147], the choice of methods is often dictated by the data type. For example, for tabular and low-dimensional data, applications of rule- and score-based methods, sparse linear models, decision trees, and k-nearest neighbours are abundant. By contrast, computer vision and natural language problems, e. g. analysis of medical images and unstructured electronic medical records, often require a different toolset. For these data types, attribution and variable importance methods highlighting relevant input regions have become increasingly popular [147]. Based on the survey among clinicians, Tonekaboni *et al.* [148] identify explanations tailored to temporal data as an underexplored research area relevant to clinical end-use. These trends confirm the general point argued by Rudin [11] that interpretability is domain-specific and its all-purpose definition is impossible.

Beyond being rich in heterogeneous data types, the healthcare domain has *multiple* stakeholder groups [34]: model developers, medical researchers, regulators, clinicians, patients, etc. Naturally, these groups may raise unique questions regarding the predictive model, its development, and its behaviour, requiring a different type of explanation or model class. Understanding how these different interest groups may influence model development, deployment, and the choice of interpretable model class or explanation technique is essential to conducting meaningful research at the interface of ML and healthcare and biomedical applications.

Generally, interpretable models and explanation methods also need to be tailored to the modelling problem at hand, and the biomedical and healthcare domains, due to their breadth, feature many distinct tasks. Holzinger *et al.* [28] observe that historically, interpretable and explainable ML have mainly focused on neural networks and supervised learning, specifically classification, trying to elucidate complex input-output relationships. However, many unexplored challenges lie beyond this limited setting, e. g. clustering, reinforcement learning, and generative modelling. Effectively translating interpretable models and explanation techniques into these scenarios remains an open problem for general machine learning and the biomedical and healthcare application domain.

Naturally, some researchers have rightfully scrutinised the utility of contemporary interpretable models and explanation methods in medical applications. For example, Ghassemi, Oakden-Rayner, and Beam [149] argue that *post hoc* explanations suffer from the "*interpretability gap*": the explanations are often not faithful to the model being explained (a similar argument is voiced by Rudin [11]) and are not necessarily helpful for understanding if the model's prediction is sensible. A similar point is demonstrated by Adebayo *et al.* [131] regarding the use of explanations to discover spurious correlations wrongly utilised by the model. While the concerns above are not unjustified, as evidenced by the content of the previous sections, interpretable and explainable ML have much to offer beyond widely criticised saliency maps and attribution measures. Nevertheless, being cautious and sceptical when examining explanations, as suggested by Ghassemi, Oakden-Rayner, and Beam [149], is a prudent course of action.

As discussed above, interpretable and explainable ML for healthcare applications must be contextualised in the type of data, e. g. time series or medical images, and modelling problem at hand, e. g. survival or medical image analysis. The breadth of data types and tasks in the different subfields of this domain necessitates the development of specialised model classes and techniques. Beyond these technical considerations, healthcare encompasses multiple stakeholder groups with varying needs for and visions of interpretability or explainability. Lastly, healthcare is undoubtedly a high-stakes domain where predictions and decisions are socially consequential. Therefore, particular caution and thought must be given to the way explanations are communicated to and treated by the end user.

2.6 SUMMARY

This chapter has provided a scoping overview of the selected methods from interpretable and explainable machine learning. The definitions and notation outlined herein will be utilised throughout the following chapters. Above, we have discussed the delineation between *ante hoc* interpretability embedded into the model class by design and *post hoc* explanation approaches applied to already-trained opaque models. As illustrated by concrete examples, the contemporary literature features a whole "zoo" of model classes and method families, reflecting different definitions and desiderata the researchers have posed for interpretability and explainability. The sheer breadth of contexts and application domains necessitates developing novel, specially tailored techniques in areas such as biomedicine and healthcare, wherein, as discussed before, unique data types and modelling problems, diverse stakeholder groups, and high-stakes decisions should be factored into the methods research.

Part I

ANTE HOC INTERPRETABLE MODELS

INTERMEZZO: ANTE HOC INTERPRETABLE MODELS

Having reviewed recent literature on interpretable and explainable ML in Chapter 2, we now turn to the first part of this thesis, tackling the design of interpretable neural network architectures (Question 1, Chapter 1; Section 2.3.4). Throughout Chapters 3, 4, and 5, we will consider different model designs tailored towards specific application scenarios. In particular, Chapter 3 will present a specialised version of self-explaining neural networks to perform time series analysis and interpret relationships between covariates observed over time. In Chapter 4, we will consider the problem of survival analysis and introduce a generative probabilistic approach for mixture modelling survival data, which facilitates subgroup discovery and prototype-based explanation of the nonlinear relationships between the covariates and survival time. Lastly, Chapter 5 will scale and generalise concept bottleneck models to the challenges pertinent to typical medical image analysis settings: the presence of multiple views and incompletely observed concept variables. In addition to various model designs and input data types, we will leverage interpretable machine learning to attain different goals, including exploratory data analysis (Chapters 3 and 4) and decision support (Chapter 5).

3

NONLINEAR TIME SERIES STRUCTURE LEARNING

As discussed in Section 2.5, unique data types are one of the salient challenges of biomedical and healthcare applications. Tonekaboni *et al.* [148] mention temporal explanations as a research question of special significance and interest. In biomedicine and healthcare, datasets are typically recorded over a period of time, with a series of data points corresponding to the same measurement unit in violation of the conventional assumption that the instances are independent and identically distributed (IID). Such sequential data, also referred to as time series or longitudinal data, require specialised modelling approaches [30], [150] and, hence, nontrivial solutions from interpretable machine learning [148]. A concrete example from the healthcare domain is the so-called patient trajectories [151], comprising records of the patient's health-related events throughout time, often stored as electronic health records (EHR) or measured using wearable sensors or other sensing technologies. Analysis of patient trajectories facilitates a better understanding of the disease dynamics and progression patterns, helpful for forecasting the patient's future course and potential adverse events.

In this chapter, we will concentrate on tackling the *exploratory analysis* of time series data to understand better the captured dynamics and relationships among the observed variables. As stated above, such analysis can be instrumental for scientific discovery and exploration of temporal patterns prevalent in the data. In particular, we will consider the inference of Granger causality (GC) [152], a practical and popular approach to the analysis of multivariate longitudinal data, named after econometrician Clive W. J. Granger who introduced this framework in his seminal work. Specifically, we will investigate the problem in the context of time series with *nonlinear* dynamics, wherein conventional statistical time series analysis methods [30] may not be as helpful.

In the following sections, we will briefly introduce the concepts of Granger causality and self-explaining neural networks (mentioned in Section 2.3.4). Subsequently, we will present a novel *interpretable* framework for inferring GC in nonlinear multivariate time series. Lastly, we will explore experimental results and describe a few concrete biomedical application examples for time series structure learning. This chapter is mainly based on the contents and text of the published work "Interpretable Models for Granger Causality Using Self-explaining Neural Networks" [153].

3.1 BACKGROUND

Throughout this chapter, let us assume being given a multivariate time series with p variables denoted by $\{\boldsymbol{x}_t\}_t = \left\{ \begin{pmatrix} x_t^1 & x_t^2 & \cdots & x_t^p \end{pmatrix}^\top \right\}_t$. Intuitively, in structure learning, we are interested in inferring pairwise temporal relationships to understand which variables "drive" each other. Such a problem is frequently tackled using relational [154], Granger-causal [152], and more general causal [155] inference techniques. In this section, we will provide essential background on Granger causality and self-explaining neural networks, which form the basis of our method introduced in Section 3.2.

3.1.1 *Granger Causality*

Granger causality [30], [152], [155] defines a set of directed pairwise relationships on a given multivariate time series. It builds on the assumption that the cause should temporally precede its effect [156]. Informally, one variable is said to *Granger-cause* another if the past values of the former are useful for predicting the future values of the latter given all other relevant information [155]. Below, we provide a more formal definition due to Tank *et al.* [95].

Definition 3.1.1 [Granger Causality, Tank *et al.* [95]]**.** Assume that the dynamics in the multivariate time series $\{\boldsymbol{x}_t\}_t$ is given by

$$x_t^i := g_i\left(x_{<t}^1, \ldots, x_{<t}^j, \ldots, x_{<t}^p\right) + \varepsilon_t^i, \text{ for } 1 \leq i \leq p, \tag{3.1}$$

where g_i, for $1 \leq i \leq p$, are (potentially nonlinear) functions, $x_{<t}^j$ denotes the past values of the j-th variable up to and including step $t-1$, i.e. $x_1^j, x_2^j, \ldots, x_{t-1}^j$, and ε_t^i are additive noise terms. We say that variable x^j *does not* Granger-cause variable x^i, denoted by $x^j \not\rightarrow x^i$, if and only if the function g^i is constant in $x_{<t}^j$.

Observe that Definition 3.1.1 generalises Granger causality beyond simple cases with linear dynamics and two variables [30], [152] since functions g_i may be nonlinear and depend on multiple covariates. Based on the definition above, pairwise Granger-causal relationships among the time series variables can be summarised as a directed graph, also referred to as a *summary graph* [155]. The Granger-causal summary graph is given by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, p\}$ and $\mathcal{E} = \left\{(i, j) : x^i \rightarrow x^j, 1 \leq i, j \leq p\right\}$, i. e. every vertex corresponds to a single variable and every directed edge indicates a Granger-causal relationship. The graph can be alternatively represented by its adjacency matrix $\boldsymbol{A} \in \{0,1\}^{p \times p}$, where $A_{i,j} = \mathbf{1}_{\{(i,j) \in \mathcal{E}\}}$. Thus, the inference of Granger causality from observational data $\{\boldsymbol{x}_t\}_{t=1}^T$ effectively amounts to learning the structure of the graph $\mathcal{G}$ and estimation of the adjacency matrix $\boldsymbol{A}$.

A conventional approach to infer Granger causality in multivariate time series with linear dynamics is to perform statistical tests on the vector autoregressive (VAR) model [30], [155], [157] fitted on the data:

$$\boldsymbol{x}_t = \boldsymbol{\nu} + \sum_{k=1}^{K} \boldsymbol{\Phi}_k \boldsymbol{x}_{t-k} + \boldsymbol{\varepsilon}_t, \tag{3.2}$$

where $\boldsymbol{\nu} \in \mathbb{R}^p$ is the intercept term, $K \geq 1$ is the order of autoregressive relationships, $\boldsymbol{\Phi}_k \in \mathbb{R}^{p \times p}$ are coefficient matrices, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon}\right)$ is the additive noise term. Under Equation 3.2, GC can be inferred by fitting a VAR model using multivariate least squares and performing pairwise statistical tests on the coefficient matrices. Namely, we conclude that $x^i \rightarrow x^j$ if, for some $k \in \{1, \ldots, K\}$, $\left(\boldsymbol{\Phi}_k\right)_{j,i} \neq 0$.

An alternative to pairwise statistical testing is to fit a regularised VAR model with sparsity-inducing regularisation applied to the coefficient matrices $\boldsymbol{\Phi}_k$ [157], [158]. An example of such an approach is the Lasso Granger method [157], which utilises a Lasso-like penalty (Section 2.3.2). This approach is particularly practical for high-dimensional time series, i. e. where p is large, for which pairwise tests might be prohibitively computationally costly or underpowered.

As mentioned, the VAR model (Equation 3.2) assumes that the observed time series have linear dynamics, i. e. all functions g_i in Equation 3.1 are linear in covariates. Many techniques have been developed to model *nonlinear* relationships. Early approaches include dynamic Bayesian networks [159] and time-smoothed logistic regression with time-varying coefficients [160]. Another family of techniques leverages kernel-based methods [161], [162]. Finally, there exists a plethora of neural-network-based approaches to GC inference [95], [96], [101], [163]–[166]. Specifically relevant to the current chapter are sparse neural networks [95], [96] and the methods based on the attention mechanism [101].

3.1.2 *Self-explaining Neural Networks*

We have briefly introduced self-explaining neural networks (SENN) [84] as one of the interpretable model classes in Section 2.3.4. Below, we will provide a more formal definition following the original work by Alvarez Melis and Jaakkola [84]. In this subsection, we will deviate from the time series setting outlined before and assume being given *static* data in the supervised learning scenario in tuples $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$.

Definition 3.1.2 [Self-explaining Neural Networks, Alvarez Melis and Jaakkola [84]]**.** A self-explaining neural network is given by

$$f\left(\boldsymbol{x}\right) = g\left(\theta\left(\boldsymbol{x}\right)_1 h\left(\boldsymbol{x}\right)_1, \ldots, \theta\left(\boldsymbol{x}\right)_k h\left(\boldsymbol{x}\right)_k\right), \tag{3.3}$$

where $g : \mathbb{R}^k \rightarrow \mathbb{R}$ is the link function and $h : \mathbb{R}^p \rightarrow \mathbb{R}^k$ maps the inputs to k interpretable basis concept variables. We refer to $\theta(\boldsymbol{x})$ as *generalised coefficients* for the data point $\boldsymbol{x}$.

In Equation 3.3, under some further assumptions described below, generalised coefficients could be utilised to explain the contributions of individual concepts to the model's output. In the simplest special case of Equation 3.3, the link function is the summation, and the concept map h is identity, i. e. individual raw features directly serve as the concepts:

$$f\left(\boldsymbol{x}\right) = \sum_{j=1}^{p} \theta\left(\boldsymbol{x}\right)_j x_j. \tag{3.4}$$

In practice, θ in Equations 3.3–3.4 is parameterised by a neural network. Also, note the resemblance between Equation 3.4 and the varying-coefficient models described in Definition 2.3.7, Section 2.3.3.

Alvarez Melis and Jaakkola [84] outline a few desiderata towards functions g, θ, and h to improve the interpretability of SENNs. (i) The link function g should be monotonic and additively separable in its inputs $z_i = \theta\left(\boldsymbol{x}\right)_i h\left(\boldsymbol{x}\right)_i$. Moreover, $\frac{\partial g}{\partial z_i} > 0$ for all $1 \leq i \leq k$. (ii) Generalised coefficients θ should be locally difference-bounded by the concept map h, i. e. for every $\boldsymbol{x}_0$, there should exist $\delta > 0$ and $L \in \mathbb{R}$ s.t. if $\left\|\boldsymbol{x} - \boldsymbol{x}_0\right\| < \delta$, then $\left\|\theta\left(\boldsymbol{x}\right) - \theta\left(\boldsymbol{x}_0\right)\right\| \leq L\left\|h\left(\boldsymbol{x}\right) - h\left(\boldsymbol{x}_0\right)\right\|$. (iii) Finally, the concepts $\{h(\boldsymbol{x})_i\}_{i=1}^{k}$ should be interpretable and representative of features $\boldsymbol{x}$, with k ideally being small. These desiderata ensure that the link function and concept variables are interpretable and the generalised coefficients are slowly varying.

To train SENNs, Alvarez Melis and Jaakkola [84] introduce a gradient-regularised loss function:

$$\ell^{y}\left(f\left(\boldsymbol{x}\right),y\right)+\lambda\ell^{\theta}\left(f\left(\boldsymbol{x}\right)\right),\tag{3.5}$$

where ℓ^{y} is the prediction loss, e. g. MSE or CE, $\lambda > 0$ is the regularisation parameter, and ℓ^{θ} is the gradient-based regulariser given by $\ell^{\theta}\left(f\left(\boldsymbol{x}\right)\right) = \left\|\nabla_{\boldsymbol{x}}f\left(\boldsymbol{x}\right)-\theta\left(\boldsymbol{x}\right)^{\top}\boldsymbol{J}_{\boldsymbol{x}}^{h}\left(\boldsymbol{x}\right)\right\|_{2}$. Note that above, $\boldsymbol{J}_{\boldsymbol{x}}^{h}$ denotes the Jacobian matrix of h w.r.t. $\boldsymbol{x}$. Thus, the gradient penalty forces f to be locally linear and allows interpreting $\theta(\boldsymbol{x})$ as local linear regression coefficients. The loss function in Equation 3.5 embodies the tradeoff between predictive performance and interpretability.

3.2 GENERALISED VECTOR AUTOREGRESSIVE MODELS

We now turn back to the time series setting to introduce an interpretable autoregressive model inspired by SENNs. Definition 3.1.1 focuses on the presence or absence of a Granger-causal relationship among variables in a nonlinear multivariate time series. Most neural-network-based frameworks for nonlinear GC, e. g. by Tank *et al.* [95], Khanna and Tan [96], and Nauta, Bucur, and Seifert [101], similarly only tackle relational inference. However, in addition to causality, practitioners might be interested in understanding the *form* of nonlinear relationships, i. e. *how* certain covariates influence each other. Some nonlinear interactions might be exclusively positive or negative—such effects are prevalent in real-world systems, for instance, consider inhibitory effects in gene regulatory networks [167] or synthesis of metabolites in metabolomics [168]. Thus, there is a need for a nonlinear GC inference framework that would allow the identification of such negative and positive interactions and would be more interpretable than relational inference.

Following the setting outlined in Definition 3.1.1, we say that variable x^{j} has a *positive* Granger-causal effect on x^{i} if the function g_i is increasing in $x^{j}_{<t}$. If g_i is decreasing in $x^{j}_{<t}$, then x^{j} has a *negative* Granger-causal effect on x^{i}. Figure 3.1 shows a schematic example of a multivariate time series with five variables alongside its summary graph. Note that the edges in the graph are coloured according to the sign of the effect. Lastly, adhering to the description above, some GC effects might be neither negative nor positive; for example, when past values of x^{j} contribute both positive and negative nonlinear terms to the function g_i at different lags.

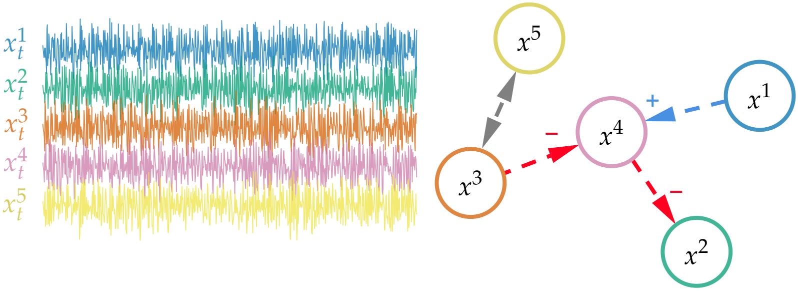


FIGURE 3.1: A schematic example of a multivariate time series (*left*) and its Granger-causal summary graph (*right*). Coloured edges correspond to **negative** and **positive** effects.

To tackle the inference of the Granger-causal effect signs under nonlinear dynamics, we introduce a novel autoregressive time series model based on self-explaining neural networks [84] and the classical vector autoregression [30] (Section 3.1). We refer to this model as *generalised vector autoregression* (GVAR).[1] Below, we provide a detailed definition and explain the main design choices.

Following the interpretable forms of the SENN and VAR models, the generalised vector autoregressive model of order K is given by

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{\Phi}_{\boldsymbol{\theta}_k}\left(\boldsymbol{x}_{t-k}\right)\boldsymbol{x}_{t-k} + \boldsymbol{\varepsilon}_t, \tag{3.6}$$

where $\boldsymbol{\Phi}_{\boldsymbol{\theta}_k} : \mathbb{R}^p \rightarrow \mathbb{R}^{p\times p}$ is a neural network with parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\varepsilon}_t$ is an additive noise term. Note that in the equation above and subsequently, we omit the intercept for brevity. The GVAR model is summarised graphically in Figure 3.2. Observe that Equation 3.6 resembles the classical VAR model (Equation 3.2). However, instead of using fixed coefficient matrices $\{\boldsymbol{\Phi}_k\}_{k=1}^{K}$, here, weights are output by K neural networks $\left\{\boldsymbol{\Phi}_{\boldsymbol{\theta}_k}\right\}_{k=1}^{K}$ and may thus be time-varying. We will refer to $\boldsymbol{\Phi}_{\boldsymbol{\theta}_k}\left(\boldsymbol{x}_{t-k}\right)$ as the generalised coefficient matrix at time step t, lag k. Observe that $\left(\boldsymbol{\Phi}_{\boldsymbol{\theta}_k}\left(\boldsymbol{x}_{t-k}\right)\right)_{i,j}$ quantifies the influence of x_{t-k}^{j} on x_t^i. Finally, note the similarity between Equations 3.6 and 3.4. The GVAR formulation above can be thought of as a self-explaining neural network with inputs $\boldsymbol{x}_{t-K}, \ldots, \boldsymbol{x}_{t-1}$ and generalised coefficients $\boldsymbol{\Phi}_{\boldsymbol{\theta}_k}$.

---

[1] Not to be confused with the global VAR by Pesaran, Schuermann, and Weiner [169].

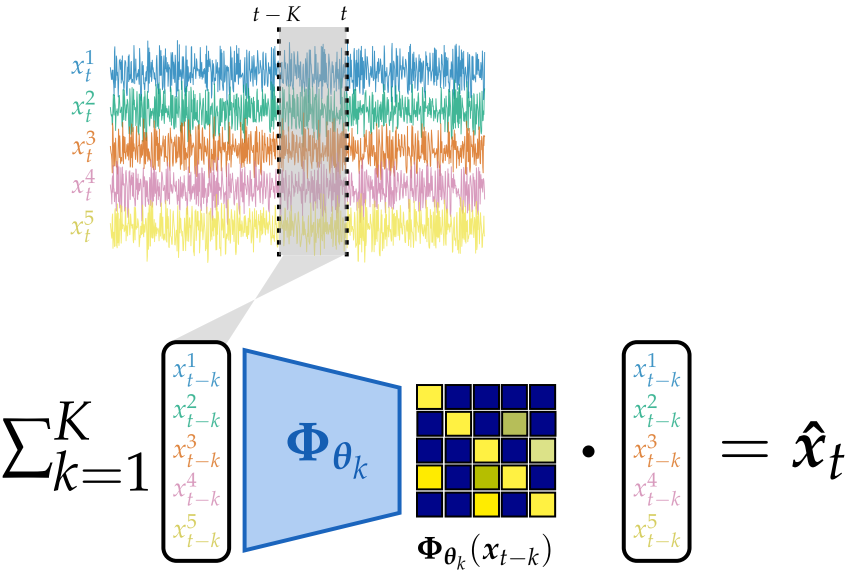


Figure 3.2: A schematic summary of the proposed generalised vector autoregressive (GVAR) model applied to a time series in five variables. Lagged time series values are fed into neural networks $\{\Phi_{\theta_k}\}_{k=1}^{K}$ to produce generalised coefficient matrices $\Phi_{\theta_k}(x_{t-k})$, which are then multiplied with the time series values. The forecast $\hat{x}_t$ is given by the sum across K lags.

Similar to SENNs (Equation 3.5) and the Lasso Granger method [157], rather than using the regular multivariate least squares objective, we resort to the regularised loss function explained below. In particular, assuming a sparse GC summary graph, we impose several penalties on the generalised coefficient matrices. Let $\Phi_t \in \mathbb{R}^{p \times Kp}$ be a shorthand notation for the concatenation of the generalised coefficient matrices obtained across all lags at time step t, i. e. $\Phi_t = \begin{bmatrix} \Phi_{\theta_K}(x_{t-K}) & \cdots & \Phi_{\theta_1}(x_{t-1}) \end{bmatrix}$. Then, given a single observed replicate of time series $\{x_t\}_{t=1}^{T}$, the GVAR is optimised by solving

$$\min_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K} \left\{ \underbrace{\frac{1}{T-K}\sum_{t=K+1}^{T}\left\|\boldsymbol{x}_t-\hat{\boldsymbol{x}}_t\right\|_2^2}_{\text{❶}} + \underbrace{\frac{\lambda}{T-K}\sum_{t=K+1}^{T}\Omega\left(\boldsymbol{\Phi}_t\right)}_{\text{❷}} \right. \\ \left. + \underbrace{\frac{\gamma}{T-K-1}\sum_{t=K+1}^{T-1}\left\|\boldsymbol{\Phi}_{t+1}-\boldsymbol{\Phi}_t\right\|_2^2}_{\text{❸}} \right\}, \tag{3.7}$$

where $\hat{\boldsymbol{x}}_t = \sum_{k=1}^{K} \boldsymbol{\Phi}_{\boldsymbol{\theta}_k}\left(\boldsymbol{x}_{t-k}\right)\boldsymbol{x}_{t-k}$ is the one-step forecast for $\boldsymbol{x}_t$ (Equation 3.6) and $\lambda, \gamma \geq 0$ are regularisation parameters.

The objective above contains three terms, which we shortly describe below. ❶ The first term is the forecasting multivariate MSE loss, which, similar to the prediction loss in SENNs (Equation 3.5), forces the model to make accurate forecasts. In the case of categorical or mixed-type time series, it should be adjusted accordingly to the CE or a weighted combination of the MSE and CE. ❷ Weighted by λ is the sparsity-inducing penalty $\Omega\left(\boldsymbol{\Phi}_t\right)$ applied to the generalised coefficient matrices across all time steps. The regulariser should be chosen following the assumptions about the sparsity of the time series GC structure. For concrete examples applicable to both the VAR and GVAR models, refer to the work by Nicholson, Matteson, and Bien [158]. For proof-of-concept experiments, we utilise an elastic-net-style penalty [69] $\Omega\left(\boldsymbol{\Phi}_t\right) = \alpha\left\|\boldsymbol{\Phi}_t\right\|_1 + (1-\alpha)\left\|\boldsymbol{\Phi}_t\right\|_2^2$ under $\alpha = 0.5$. ❸ Finally, the last term weighted by γ is the smoothing penalty, which corresponds to the mean squared difference between generalised coefficient matrices across consecutive time steps. This term enforces smoothness in generalised coefficients throughout time and, akin to the SENN gradient penalty in Equation 3.5, has a linearising effect on the model, e. g. in a linear VAR, since the coefficients are not time-varying, $\left\|\boldsymbol{\Phi}_{t+1}-\boldsymbol{\Phi}_t\right\|_2^2 = 0$.

In summary, the GVAR model is optimised by jointly training K neural networks to predict generalised coefficient matrices. Its loss function is regularised to encourage sparse and slowly varying coefficients, reflecting the assumptions about the sparsity of the Granger-causal summary graph and the linearity of the autoregressive relationships.

3.2.1 *Inferring Granger Causality*

As explained before, generalised coefficient matrices quantify the contributions of the lagged time series values to the forecast (Equation 3.6, Figure 3.2). By inspecting the magnitudes, signs, and variability of these coefficients, we can infer Grager-causal relationships among time series variables, including the effect signs (Figure 3.1). This subsection introduces the framework for inferring Granger causality based on the GVAR.

Given a single replicate of a multivariate time series $\{\boldsymbol{x}_t\}_{t=1}^{T}$, we first optimise the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ according to Equation 3.7. Using trained neural networks $\boldsymbol{\Phi}_{\hat{\boldsymbol{\theta}}_k}$, $1 \leq k \leq K$, we compute summary statistics for GC relationships represented by a matrix $\boldsymbol{S} \in \mathbb{R}^{p \times p}$:

$$S_{i,j} = \max_{1 \leq k \leq K} \left\{ \mathrm{median}_{K+1 \leq t \leq T} \left(\left| \left(\boldsymbol{\Phi}_{\hat{\boldsymbol{\theta}}_k} \left(\boldsymbol{x}_t \right) \right)_{i,j} \right| \right) \right\}, \text{ for } 1 \leq i, j \leq p. \quad (3.8)$$

Above, $S_{i,j}$ quantifies the strength of the GC effect of x^i on x^j. In particular, the statistic corresponds to the maximum value across lags of the median absolute generalised coefficient. Ideally, for causal variable pairs $x^i \rightarrow x^j$, we would expect that $S_{i,j} > 0$, whereas, in noncausal relationships, it should be close to 0 since, due to the sparsity-inducing regularisation in Equation 3.7, noncausal coefficients should be shrunk towards 0. Note that choices of summary statistics other than the one proposed in Equation 3.8 are viable. For instance, the median or maximum could be replaced with the average.

In practice, after training the GVAR model, the generalised coefficients are not shrunk to exact zeros, and consequently, for noncausal pairs, the summary statistics above are not exactly equal to 0. To tackle this and estimate the adjacency matrix of the GC summary graph, we introduce a heuristic *forward-backward stability-based thresholding* procedure, which alongside the summary statistics from Equation 3.8 (*forward*), leverages time-reversed Granger causality (TRGC) [170]–[172] (*backward*) by fitting the model and performing inference on the time-reversed data. Intuitively, this procedure identifies a threshold value for the summary statistics s.t. the inferred summary graph skeleton is stable across the original and time-reversed models. This procedure is based on the observation that on the time-reversed data, under some further assumptions, the direction of the Granger-causal effects is reversed [172]. This method is summarised as pseudocode in Algorithm 1.

Algorithm 1: Forward-backward Stability-based Thresholding

```
Input: A replicate of multivariate time series $\{\boldsymbol{x}_t\}_{t=1}^{T}$; regularisation
       parameters $\lambda, \gamma \geq 0$; model order $K \geq 1$; strictly increasing
       sequence $\boldsymbol{\xi} = \left(\xi_1, \ldots, \xi_Q\right)$, where $0 \leq \xi_1 < \xi_Q \leq 1$
Output: Estimate $\hat{\boldsymbol{A}}$ of the adjacency matrix of the GC summary
        graph

Let $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^{T}$ be the time-reversed series, i.e. let
  $\{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_T\} \leftarrow \{\boldsymbol{x}_T, \ldots, \boldsymbol{x}_1\}$
Let $\tau : \mathbb{R}^{p \times p} \times \mathbb{R} \rightarrow \{0, 1\}^{p \times p}$ be an elementwise thresholding
  function. For $1 \leq i, j \leq p$, let $\left(\tau\left(\boldsymbol{X}, \chi\right)\right)_{i,j} := \mathbf{1}_{\{|X_{i,j}| \geq \chi\}}$
Train an order $K$ GVAR model with the parameters $\lambda$ and $\gamma$ by
  minimising the loss function in Equation 3.7 on $\{\boldsymbol{x}_t\}_{t=1}^{T}$ and
  compute $\boldsymbol{S}$ using Equation 3.8
Train another GVAR on $\{\tilde{\boldsymbol{x}}_t\}_{t=1}^{T}$ and compute $\tilde{\boldsymbol{S}}$ using Equation 3.8
for $i = 1$ to $Q$ do
    Let $\kappa_i \leftarrow q_{\xi_i}(\boldsymbol{S})$ and $\tilde{\kappa}_i \leftarrow q_{\xi_i}(\tilde{\boldsymbol{S}})$, where $q_{\xi}(\boldsymbol{X})$ denotes the
      $\xi$-quantile of the components of $\boldsymbol{X}$
    Evaluate agreement
      $\varsigma_i \leftarrow \frac{1}{2}\left[\mathrm{BA}\left(\tau\left(\boldsymbol{S}, \kappa_i\right), \tau\left(\tilde{\boldsymbol{S}}^{\top}, \tilde{\kappa}_i\right)\right) + \mathrm{BA}\left(\tau\left(\tilde{\boldsymbol{S}}^{\top}, \tilde{\kappa}_i\right), \tau\left(\boldsymbol{S}, \kappa_i\right)\right)\right]$,
      where BA denotes the balanced accuracy
end
Let $i^* \leftarrow \arg\max_{1 \leq i \leq Q} \varsigma_i$ and $\xi^* \leftarrow \xi_{i^*}$
Let $\hat{\boldsymbol{A}} \leftarrow \tau\left(\boldsymbol{S}, q_{\xi^*}(\boldsymbol{S})\right)$

return $\hat{\boldsymbol{A}}$
```

First, GVAR models are trained on the original time series and time-reversed data and used to compute matrices with the summary statistics $\boldsymbol{S}$ and $\tilde{\boldsymbol{S}}$ (lines 3–4). Subsequently, iterating across Q threshold values κ_i and $\tilde{\kappa}_i$ given by the ξ_i-quantiles of $\boldsymbol{S}$ and $\tilde{\boldsymbol{S}}$ (lines 5–8), respectively, we evaluate the agreement ς_i between thresholded matrices $\tau\left(\boldsymbol{S}, \kappa_i\right)$ and $\tau\left(\tilde{\boldsymbol{S}}^{\top}, \tilde{\kappa}_i\right)$ by computing the averaged balanced accuracy (BA) [173] on the matrix components. Observe that the matrix $\tilde{\boldsymbol{S}}$ is transposed because we expect the GC relationships to be flipped on the time-reversed series. We utilise balanced accuracy as the agreement measure due to the assumed sparsity of the GC summary graph and the fact that the BA weighs true positives and negatives equally. However, other choices of the agreement measure are plausible, e. g. graph similarity scores [174]. To avoid inferring trivial causal structures, namely, the null graph or the graph with only self-causal links, we set the agreement of such solutions to 0. In the last steps of the algorithm, we retrieve an optimal threshold $q_{\xi^*}\left(\boldsymbol{S}\right)$ corresponding to the maximum agreement between thresholded summary statistics matrices and return the thresholded matrix $\hat{\boldsymbol{A}} = \tau\left(\boldsymbol{S}, q_{\xi^*}\left(\boldsymbol{S}\right)\right)$ as an estimate of the adjacency matrix of the GC summary graph (lines 9–10).

In summary, the forward-backward stability-based thresholding procedure introduced above tries to find a causal structure that is *stable* across the original and time-reversed series. Intuitively, it exploits a principle similar to the stability selection [175], [176], a method for selecting an appropriate sparsity level based on the error rate control in high-dimensional inference problems.

3.2.2 *Further Remarks*

Many prior neural-network-based GC inference frameworks [95], [96], [101] require training a separate neural network for every variable, i. e. p neural networks in total. In contrast, the GVAR model and forward-backward stability-based thresholding utilise $2K$ networks. Thus, compared to the previous works, GVAR could be especially useful in time series where $K \ll p$. However, this reduction in the number of models to be trained comes at the cost of memory complexity: each of the K networks of a GVAR model has p^2 outputs, and all K networks need to be optimised jointly.

One of the crucial design choices is the architecture of neural networks $\boldsymbol{\Phi}_{\boldsymbol{\theta}_k}$ in Equation 3.6. Although in the experiments presented in the following section, we utilise MLPs with p inputs and p^2 outputs, other network architectures are plausible, e. g. long short-term memory (LSTM) [177], temporal convolutional networks (TCN) [178], or attention-based architectures [179]. Using more powerful architectures and specialised inductive biases could improve the forecasting performance of the GVAR and, consequently, help causal inference.

Lastly, our framework has important hyperparameters that can influence the model's predictive performance and inference results. Specifically, the weights λ and γ of the penalty terms in the loss function (Equation 3.7) have a significant influence on the sparsity and temporal smoothness of the generalised coefficient matrices. A naïve practical approach to choosing these hyperparameters is tuning them w.r.t. the forecasting error on held-out validation data. However, models with the best predictive performance may not always be consistent with the ground-truth structure [175]. Therefore, similar to the thresholding procedure in Algorithm 1, one may resort to a stability-based approach [175] to choose appropriate hyperparameter values.

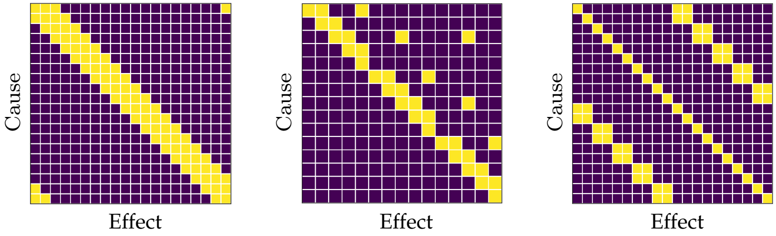
3.3 EXPERIMENTS AND RESULTS

This section describes the empirical findings from evaluating the GVAR model and inference framework introduced in Section 3.2 alongside several baseline techniques from the related literature. Our goal is to compare the methods' performance at the inference of (i) the GC structure and (ii) effect signs (Section 3.1.1). The experiments focus on several synthetic benchmarking datasets with nonlinear dynamics and known ground-truth GC summary graphs. In the following subsections, before describing the results, we briefly explain the baseline techniques and benchmarking datasets utilised for evaluation.

3.3.1 *Baseline Methods*

As the simplest baseline approach, we fit a linear VAR model (Equation 3.2) and perform a series of statistical tests for Granger causality using the Benjamini–Hochberg procedure [180] for the false discovery rate (FDR) control ($q = 0.05$). In particular, p^2 F-tests are performed with the null hypothesis of no Granger causality for each pair of variables.

We also compare the GVAR model to several neural-network-based approaches. Closely related to this work are sparse neural networks, namely, componentwise multilayer perceptron (cMLP) and long short-term memory (cLSTM) [95] and economy statistical recurrent unit (eSRU) [96]. Similar to



(a) Lorenz 96 (b) fMRI (c) Lotka–Volterra

FIGURE 3.3: Granger-causal summary graph adjacency matrices for (a) Lorenz 96, (b) fMRI, and (c) multispecies Lotka–Volterra time series. **Dark**-coloured cells denote a lack of GC relationships ($A_{i,j} = 0$), whereas **light**-coloured cells correspond to causal links ($A_{i,j} = 1$).

ours, all three approaches mentioned above can learn nonlinear dynamics and rely on regularisation to infer a sparse GC summary graph. Last, we consider the temporal causal discovery framework (TCDF) by Nauta, Bucur, and Seifert [101], which leverages the temporal convolutional networks and regularised attention mechanism to perform GC discovery.

For all experiments and methods compared, we perform a grid search across hyperparameters controlling the sparsity of inferred GC summary graphs. For instance, for the GVAR model, the search was conducted over the values of λ- and γ-parameters (Equation 3.7). For a fair comparison, the final results reported correspond to the methods' maximal performance w.r.t. BA (Section 3.3.3). Finally, if applicable, all models are fixed to the same order of autoregressive relationships. Neural network architectures utilised by the GVAR are described in Appendix A.1.

3.3.2 *Benchmarking Datasets*

As mentioned above, the evaluation is performed on *synthetic* data, which allows for validation against the known ground-truth causal structure. Figure 3.3 contains visualisations of the GC summary graphs for all three benchmarks. Below, we summarise each dataset, explaining the underlying generative process and principal hyperparameters.

A conventional benchmark from the time series structure learning literature is the Lorenz 96 model introduced by Edward N. Lorenz [181] as a toy problem to study error growth, chaotic behaviour, and predictability in weather dynamics. It is a continuous-time dynamical system comprising $p \geq 4$ nonlinear differential equations:

$$\frac{dx^i}{dt} = \left(x^{i+1} - x^{i-2}\right) x^{i-1} - x^i + F, \text{ for } 1 \leq i \leq p, \tag{3.9}$$

where, by definition, $x^0 := x^p$, $x^{-1} := x^{p-1}$, and $x^{p+1} := x^1$. Above, F is a forcing constant, which, in combination with p, affects the nonlinearity of the dynamics. As shown in Equation 3.9 and Figure 3.3a, for larger values of p, the resulting GC summary graph is sparse since every variable Granger-causes and is caused by just three other variables. We perform experiments on five independent replicates of the time series generated by numerically simulating the Lorenz 96 system. In particular, we consider data with $p = 5$ variables $T = 500$ time steps long under $F = 10$ and 40. Tank *et al.* [95] and Khanna and Tan [96] explore the same forcing constant values.

A more challenging benchmark is rich and realistic blood-oxygen-level-dependent (BOLD) time series simulations from the NetSim dataset [182]. These simulations were initially introduced to evaluate network modelling techniques for functional magnetic resonance imaging (fMRI) time series. In this context, variables represent spatial regions of interest within the human brain [182]. Thus, the Granger-causal inference allows estimating the brain "network" to explore influences among neuronal populations [183]. In this section, we perform experiments on five $T = 200$ step long replicates of the third simulation from the NetSim dataset, containing $p = 15$ variables. The ground-truth times series structure is visualised in Figure 3.3b.

Finally, to evaluate GC effect sign inference, we utilise the Lotka–Volterra model [184], extending it beyond the two-variable setting. This model, introduced independently by Alfred J. Lotka and Vito Volterra, is a toy representation of the population dynamics in ecology. The original dynamical system [184] comprises two differential equations describing the dynamics of the predator and prey populations, denoted by y and x, respectively:

$$\begin{aligned} \frac{dx}{dt} &= \alpha x - \beta xy, \\ \frac{dy}{dt} &= -\rho y + \delta yx, \end{aligned} \tag{3.10}$$

where $\alpha, \beta, \delta, \rho > 0$ are parameters controlling the strength of causal interactions. Observe that in Equation 3.10, y (predators) exerts only a *negative* influence on x (prey) via the term $-\beta xy$, whereas y has a *positive* effect on x given by δyx.

For this chapter, we extend the Lotka–Volterra model by including *multiple* predator and prey species, each interacting with a few other populations. This *multispecies* Lotka–Volterra model is given by the following system of differential equations:

$$\begin{aligned}
\frac{dx^i}{dt} &= \alpha x^i - \beta x^i \sum_{j=\lfloor (i-1)/q \rfloor q+1}^{\lfloor (i-1)/q \rfloor q+q} y^j - \eta \left(x^i\right)^2, && \text{for } 1 \leq i \leq p/2, \\
\frac{dy^i}{dt} &= -\rho y^i + \delta y^i \sum_{j=\lfloor (i-1)/q \rfloor q+1}^{\lfloor (i-1)/q \rfloor q+q} x^j, && \text{for } 1 \leq i \leq p/2,
\end{aligned} \tag{3.11}$$

where x^i and y^i denote population sizes of the prey and predator species, respectively, p is the total number of variables assumed to be even, $\alpha, \beta, \delta, \eta, \rho > 0$ are fixed parameters, and $q \geq 1$ is the number of Granger causes for each variable (excluding itself). In Equation 3.11, every prey species x^i is negatively affected by q predator species; likewise, every y^i is positively affected by q prey populations. Throughout the experiments, we set $\alpha = \rho = 1.1$, $\beta = \delta = 0.2$, $\eta = 2.75 \times 10^{-5}$, $q = 2$, and $p = 20$. We simulate the model five times numerically for $T = 2000$ steps using the Runge–Kutta method. During simulation, we clip all variable values to be nonnegative and introduce Gaussian noise. The GC summary graph under the parameter values mentioned above is shown in Figure 3.3c.

3.3.3 *Evaluation Metrics*

To compare the Granger-causal inference techniques considered (Section 3.3.1), we train the models and perform inference on every time series replicate, evaluating the inferred GC structure against the known ground truth (Figure 3.3). For the thresholded outputs, namely, the output of the forward-backward stability-based thresholding (Algorithm 1), we compute the accuracy (ACC) and balanced accuracy. Across all evaluation metrics, we ignore self-causal relationships. For cMLP, cLSTM, and eSRU, thresholding is performed by comparing relevant layer weight norms to 0, whereas TCDF uses a specially designed permutation test. Recall that the VAR model estimates the GC structure using hypothesis testing, adjusting the resulting p-values for multiple comparisons.

In addition to the binary-valued adjacency matrices, we examine the statistics that all methods utilise to infer GC. For our model, we look at the summary statistics computed for each variable pair according to Equation 3.8. For VAR, we inspect adjusted p-values from the pairwise F-tests. We consider relevant layer weight norms for cMLP, cLSTM, and eSRU, while in the TCDF, we assess the attention scores. These statistics are evaluated using areas under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves [185].

Finally, we slightly adjust the statistics above to evaluate the models' performance at the effect sign inference. In particular, we look at the coefficient signs for the VAR model and at the median generalised coefficient signs for GVAR. For cMLP, cLSTM, eSRU, and TCDF, we consider the signs of averaged weights from the relevant layers. We compute balanced accuracy separately for positive (BA_{pos}) and negative (BA_{neg}) GC relationships.

3.3.4 *Structure Learning*

To begin with, we compare GVAR and baseline techniques at inferring Granger-causal structure under nonlinear dynamics. The comparison is conducted on the Lorenz 96 and fMRI time series (Section 3.3.2) by evaluating inferred time series structures against known ground truth. Table 3.1 provides a summary of the results averaged across multiple independent replicates.

As mentioned, for the Lorenz 96 model, we consider two simulation settings under different values of the forcing constant (Equation 3.9). Under $F = 10$, most models successfully identify the presence or lack of GC relationships, including linear VAR, with the TCDF systematically underperforming and GVAR attaining the highest average performance across all metrics. For $F = 40$, the variability in performance across methods increases, maintaining a similar pattern: GVAR achieves the highest (balanced) accuracy while having the second-best AUROC and AUPR after cMLP. In this setting, GVAR has the most well-balanced performance w.r.t. both inferring the "hard" GC structure (ACC and BA) and providing summary statistics on the strength of causal relationships (AUROC and AUPR).

A relatively more challenging benchmark considered in this experiment is the synthetic fMRI BOLD time series. In contrast to Lorenz 96, TCDF is the most performant technique, beating other approaches w.r.t. BA, AUROC, and AUPR. It is followed by GVAR and cLSTM, which perform comparably across most metrics. Interestingly, eSRU trained using the proximal gradient descent algorithm [186] fails to shrink relevant weight to zeros or shrinks all of its weights, consequently erroneously inferring the full or null causal graph. For this reason, we do not report accuracy in this case, marking relevant entries by “—” in Table 3.1. Lastly, although the VAR has very high accuracy, its BA is close to the performance of the fair coin flip, suggesting that the linear model is not as helpful in inferring the structure on this benchmark and is underpowered, i. e. wrongly fails to reject the null hypothesis in the majority of pairwise tests.

Dataset	Model	ACC	BA	AUROC	AUPR
Lorenz 96, $F = 10$	VAR	0.92±0.01	0.84±0.02	0.94±0.02	0.83±0.03
	cMLP	*0.97±0.01*	*0.96±0.02*	*0.96±0.02*	0.91±0.05
	cLSTM	*0.97±0.01*	0.95±0.03	*0.96±0.03*	0.93±0.05
	TCDF	0.87±0.01	0.71±0.04	0.86±0.03	0.60±0.05
	eSRU	*0.97±0.01*	0.95±0.02	*0.96±0.02*	*0.94±0.03*
	GVAR	**0.98±0.00**	**0.98±0.01**	**1.00±0.00**	**0.98±0.02**
Lorenz 96, $F = 40$	VAR	0.86±0.01	0.59±0.03	0.75±0.05	0.47±0.04
	cMLP	0.68±0.03	*0.81±0.02*	**0.98±0.02**	**0.96±0.03**
	cLSTM	0.84±0.01	0.66±0.04	0.66±0.04	0.39±0.06
	TCDF	0.78±0.02	0.60±0.03	0.68±0.02	0.31±0.05
	eSRU	*0.87±0.01*	**0.89±0.02**	0.93±0.02	0.83±0.03
	GVAR	**0.95±0.01**	**0.89±0.05**	*0.97±0.01*	*0.92±0.02*
fMRI	VAR	**0.91±0.01**	0.51±0.02	0.62±0.04	0.18±0.05
	cMLP	0.85±0.03	0.61±0.07	0.62±0.07	0.19±0.06
	cLSTM	0.83±0.02	*0.66±0.05*	0.66±0.05	0.23±0.06
	TCDF	*0.90±0.02*	**0.73±0.06**	**0.81±0.04**	**0.37±0.13**
	eSRU	—	—	0.65±0.06	0.19±0.10
	GVAR	0.81±0.07	0.65±0.05	*0.69±0.07*	*0.29±0.12*

TABLE 3.1: GC structure learning results on the Lorenz 96 (under $F = 10$ and 40) and fMRI time series w.r.t. accuracy (ACC), balanced accuracy (BA), and areas under receiver operating characteristic (AUROC) and precision-recall (AUPR) curves. Results are reported as averages and standard deviations across five independent replicates. **Bold** indicates the best results, *italics* indicates the second best.

3.3.5 *Effect Sign Inference*

In addition to inferring the GC structure in nonlinear time series, we investigate the inference of GC effect signs (Section 3.1.1), i. e. the ability of the models to differentiate between positive and negative interactions. To this end, we perform inference on the multispecies Lotka–Volterra time series (Equation 3.11), where certain Granger-causal relationships can be labelled as positive or negative. In this experiment, we fix the order of autoregressive dependencies to $K = 1$ across all models that allow explicitly specifying this parameter.

Table 3.2 reports the accuracy for inferring the GC structure and balanced accuracy for detecting positive and negative effects, as described in Section 3.3.3. While most techniques are able to learn at least partially correct causal structure, few models can differentiate the effect signs. While the coefficients in the linear model are informative, as suggested by BA_{pos} and BA_{neg}, VAR fails to infer the correct structure. Surprisingly, cMLP achieves an overall well-balanced performance, and its input-layer weights are systematically correlated with the signs of the GC effects. By contrast, cLSTM, TCDF, and eSRU perform poorly at effect sign detection. Moreover, cLSTM, similar to eSRU, uses the proximal gradient descent and fails to shrink its relevant weights to zeros. Hence, we do not report its structure inference results. As intended, the GVAR model has the best results by a margin, especially for differentiating between positive and negative relationships.

Model	**ACC**	**BA**	**BA_{pos}**	**BA_{neg}**
VAR	0.38±0.10	0.64±0.06	0.85±0.02	0.78±0.04
cMLP	*0.83±0.04*	*0.83±0.04*	*0.89±0.03*	*0.85±0.08*
cLSTM	—	—	0.49±0.03	0.60±0.04
TCDF	*0.83±0.01*	0.50±0.01	0.54±0.05	0.50±0.09
eSRU	0.70±0.05	0.76±0.01	0.50±0.03	0.65±0.08
GVAR	**0.98±0.01**	**0.96±0.01**	**0.93±0.03**	**1.00±0.00**

TABLE 3.2: GC structure learning and effect sign inference results on the multispecies Lotka–Volterra time series. In addition to accuracy (ACC) and balanced accuracy (BA), we report balanced accuracy for positive (BA_{pos}) and negative (BA_{neg}) GC effects.

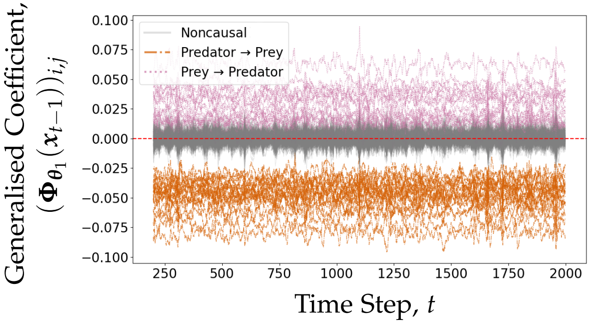
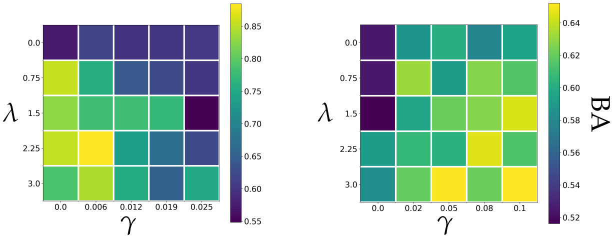


FIGURE 3.4: Plot of the generalised coefficients across time from the GVAR model trained on the multispecies Lotka–Volterra time series. Every trace corresponds to a single generalised coefficient $(\boldsymbol{\Phi}_{\boldsymbol{\theta}_1}(\boldsymbol{x}_{t-1}))_{i,j}$, for $1 \leq i \neq j \leq p$. Traces are coloured according to the ground-truth GC structure and effect signs: **grey** shows coefficients for noncausal relationships, **orange** shows coefficients for the negative predator → prey relationships, **pink** shows coefficients for the positive prey → predator relationships.

Figure 3.4 shows the variability in the generalised coefficients of the GVAR model for the multispecies Lotka–Volterra time series. In particular, we plot the coefficients $\left(\boldsymbol{\Phi}_{\boldsymbol{\theta}_1}(\boldsymbol{x}_{t-1})\right)_{i,j}$, for $1 \leq i \neq j \leq p$, across time t. We observe that while the generalised coefficients vary with time, their signs are correlated with ground truth. For the effects of the predator on prey species, the corresponding generalised coefficients are consistently negatively valued across all time steps. Similarly, the coefficients are nearly always above zero for the effects of prey populations on predators. Finally, the coefficients of noncausal variable pairs fluctuate around zero and have low magnitudes. Thus, in agreement with the results reported in Table 3.2, Figure 3.4 shows that GVAR's coefficients are, indeed, helpful for determining effect signs during GC inference.

3.3.6 *Ablation Analysis of the Loss Function*

To explore the behaviour of the GVAR model under varying hyperparameter values, we perform an ablation study on its loss function (Equation 3.7). Specifically, we investigate the role of the sparsity-inducing and smoothing penalty terms weighted by the parameters λ and γ, respectively. We run a



(a) Lorenz 96

(b) fMRI

Figure 3.5: Results of the ablation study on the GVAR's loss function for the (a) Lorenz 96 ($F = 40$) and (b) fMRI datasets. A grid search was performed across values of λ and γ, which weigh sparsity-inducing and smoothing penalties, respectively. Every cell shows the averaged balanced accuracy (BA) for a single hyperparameter configuration. **Darker** colours correspond to lower BA, and **lighter** colours denote higher BA.

grid search across several values of these parameters and report changes in the accuracy of the inferred GC structure.

Figure 3.5 displays grid search results on the Lorenz 96 and fMRI time series. For both benchmarks, regularisation is instrumental for accurately inferring the GC summary graph since, when both λ and γ are set to zeros, the performance of GVAR is close to the balanced accuracy of 0.5. Similar patterns are observed for other metrics, omitted for brevity. For Lorenz 96 (Figure 3.5a), the sparsity-inducing penalty is especially important, whereas higher values of γ affect the performance adversely. By contrast, on fMRI (Figure 3.5b), the highest accuracy is achieved when both penalties are weighted highly. These observations suggest that accurate GC inference requires careful fine-tuning of the two loss parameters.

3.4 DISCUSSION

Below, we summarise this chapter, recapitulating our main contributions to the literature and pinpointing this work in the broader context of the current doctoral thesis. Subsequently, we provide a comprehensive discussion of the empirical findings described in the previous section. Additionally, we outline several potential application examples from the biomedical domain for the proposed method. Lastly, we reflect on this work's limitations and promising future directions.

This chapter has explored structure learning in multivariate time series with nonlinear dynamics, treating this problem from the perspective of Granger-causal inference [152]. Specifically, we have developed an *interpretable* neural-network-based time series model by combining the conventional vector autoregression [30] with self-explaining neural networks [84]. The resulting generalised vector autoregression uses neural networks to learn time-varying generalised coefficients, which are regularised to be sparse and vary smoothly with time. Furthermore, we introduce the forward-backward stability-based thresholding procedure, which infers the time series structure using these generalised coefficients and time-reversed GC [170]. Since the time-varying coefficients are regularised, (i) the inferred structure is *sparse* with few GC interactions, and (ii) the form of GC relationships can be interpreted by inspecting the signs, magnitudes, and variability of the corresponding coefficients. We demonstrate the efficacy of the proposed time series model and inference technique in experiments on synthetic data with known causal structures.

The GVAR model (Equation 3.6) extends SENNs to the autoregressive setting in several ways. Firstly, instead of utilising a single network to map features to a vector of varying coefficients, we jointly train *several* networks to extract coefficient *matrices*. Furthermore, we replace the gradient penalty in the vanilla loss function (Equation 3.5) with sparsity-inducing and smoothing regularisers (Equation 3.7) that reflect typical assumptions on the GC structure of the time series: (i) the causal summary graph is sparse, and (ii) the relationships between variables evolve slowly over time.

Several salient characteristics set our method apart from the previous related literature on neural-network-based approaches to GC discovery. Componentwise MLP and LSTM [95] and eSRU [96] rely on inducing sparsity in certain layers of neural networks and use weight norms to infer GC relationships. The TCDF [101] uses attention-based scores to assess causal interactions between variables. By contrast, GVAR builds on the simple VAR model, augmenting it with varying coefficients that are regularised for further interpretability. As mentioned, this approach is more amenable to exploratory analysis and, for instance, allows the identification of positive and negative GC effects. The most closely related concurrent work is neural additive vector autoregression (NAVAR) by Bussmann, Nys, and Latré [187]. This model class leverages additivity (Section 2.3.3) w.r.t. individual variables and sparsity to improve the interpretability of GC discovery. Observe that, in contrast, GVAR generally does not assume additive relationships.

Another differentiating aspect of our inference framework is the computational complexity (Section 3.2.2): while cMLP, cLSTM, eSRU, and TCDF require training p neural networks, one modelling every variable in the time series, our forward-backward stability-based thresholding procedure trains $2K$ neural networks. Thus, our approach can lead to a significantly lower training time in high-dimensional time series with low-order autoregressive dependencies, i. e. under $K \ll p$.

The proposed method is also related to several conventional statistical models with time-varying coefficients [160], [188], [189], specialised variants of VAR in particular. Although such models allow representing complex time-varying autoregressive relationships, they rely on stricter assumptions about the evolution of coefficients and require specialised and computationally expensive estimation routines, e. g. Bayesian Markov chain Monte Carlo (MCMC) methods. Using neural networks allows GVAR to simplify and speed up the fitting without making restrictive assumptions.

Finally, another technical contribution of this chapter is using time-reversed GC and stability in the context of nonlinear GC inference. While TRGC has been studied extensively in the preceding literature [170]–[172], our work is among the first to incorporate this concept into a neural-network-based framework. The resulting forward-backward stability-based thresholding procedure allows inferring the adjacency matrix of the GC summary graph. Other related techniques rely either on sparsity penalties and proximal gradient descent [95], [96] or statistical testing [101].

We now turn to the empirical findings presented in the current chapter. We have performed a comprehensive evaluation of the GVAR model and stability-based thresholding alongside relevant baseline techniques. The evaluation was conducted on several synthetic time series datasets (Section 3.3.2) with known ground-truth GC structures (Figure 3.3). Furthermore, to demonstrate the interpretability of GVAR, we applied it to the GC effect sign inference on the specially designed multispecies Lotka–Volterra model (Equation 3.11), which features exclusively positive and negative effects among some variables.

On the Lorenz 96 (Equation 3.9) and fMRI time series, we observed that GVAR attained overall well-balanced performance at structure learning across all metrics considered (Table 3.1). While our method achieved superior results compared to baselines under two different forcing constant values on Lorenz 96, for the fMRI time series, TCDF considerably outperformed our and other techniques. Nevertheless, the performance of TCDF was not satisfactory on several other benchmarks (Tables 3.1–3.2). Our structure learning experiments also demonstrated that the proximal gradient descent utilised by cMLP, cLSTM, and eSRU to train sparse neural networks is not always effective and sometimes fails to shrink any weights to exact zeros. In general, we conclude that GVAR combined with forward-backward stability-based thresholding performs comparably to other neural-network-based approaches for time series GC structure learning under nonlinear dynamics.

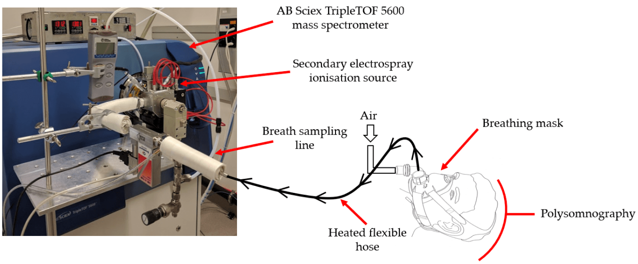
For the effect sign inference on the Lotka–Volterra model, GVAR considerably outperformed other approaches (Table 3.2), attaining near-perfect accuracy. This finding matches our expectations, as the GVAR model is, by design, interpretable and can be leveraged to explore GC relationships via generalised coefficients that are sparse and slowly varying (Figure 3.4). By contrast, in most other neural networks, the statistics we defined for GC inference were not correlated with effect signs, except for cMLP, which performed better than the linear VAR model. Thus, our method is competitive at structure inference and more interpretable than sparse neural networks and attention-based models.

Lastly, we explored the GVAR's loss function (Equation 3.7) in an ablation experiment by varying hyperparameters λ and γ (Figure 3.5) corresponding to the weights of sparsity-inducing and smoothing penalties. The results suggest that both regularisation terms are crucial for accurately inferring the GC structure. Therefore, the introduction of both regularisers is justified by our empirical findings. In practice, both parameters would need to be carefully fine-tuned, for instance, on held-out data or following a procedure similar to the stability selection [175].

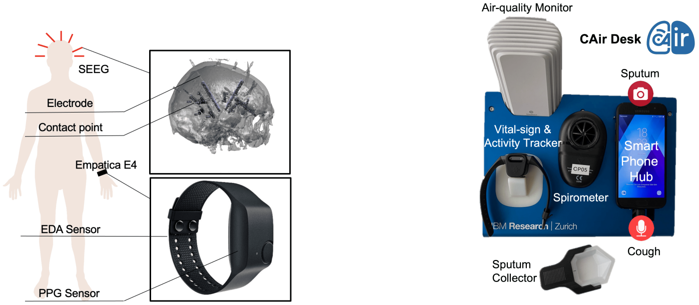
3.4.1 *Potential Applications*

The findings described above focus entirely on the synthetic datasets, the evaluation on which is endemic to time series structure learning literature, given the lack of real-world benchmarks with known ground-truth causal relationships. This section provides concrete examples of the potential use cases for nonlinear time series structure learning methods from the biomedical application domain. Figure 3.6 contains schematic summaries of the experimental setups of the three application scenarios outlined below.

The human organism can be thought of as a network of complex and continuously interacting physiological systems [190]. Thus, inferring interactions from longitudinal observations can help understand fundamental relationships and processes within the human body. One example is



(a) Breathomics and sleep stage data analysis to understand the relationship between human sleep and metabolism. Taken from Marcinkevics [191].



(b) Intracranial stereoelectroencephalography (SEEG), photoplethysmography (PPG), and electrodermal activity (EDA) analysis. Taken from Hatteland *et al.* [193]. © 2021 IEEE.

(c) Multisensory home-device and self-reported symptoms data analysis in chronic obstructive pulmonary disease (COPD). Taken from Xiao *et al.* [194].

Figure 3.6: Examples of potential biomedical applications of nonlinear time series structure learning from the related literature. Structure learning can uncover temporal relationships between biosignals to (a, b) help understand fundamental interactions in human physiology and (c) identify data modalities relevant to forecasting patients' outcomes.

the relationship between human metabolism and sleep [191], [192] (Figure 3.6a), which is poorly understood and can be elucidated using analytical techniques from chemistry, medical diagnostic tools, and time series analysis methods, such as the one introduced in this chapter. In particular, Nowak *et al.* [192] leverage neural-network-based GC inference techniques to analyse autoregressive dependencies between sleep stage transitions derived from polysomnography and metabolic activity given by volatile organic compound abundance in the exhaled breath measured using high-frequency mass spectrometry. The latter analysis is often referred to as *breath metabolomics* or *breathomics* [195]. In this context, GVAR and forward-backward stability-based thresholding can help infer sparse and interpretable relationships between sleep states and numerous metabolites, allowing us to hypothesise which volatile organic compound abundances are driven up or down during specific stages of sleep.

Another related application scenario is the exploration of the relationships between peripheral and brain biosignals [193] (Figure 3.6b). Similar to the use case above, many uncertainties surround relationships between autonomic peripheral activity and brain autonomic centres. To this end, Hatteland *et al.* [193] analyse associations between electrodermal activity, photoplethysmographic signals, being proxies for sweat response and heart rate, respectively, and stereoelectroencephalography (SEEG) data, comprising encephalographic (EEG) signals recorded from electrodes implanted into the brain. Specifically, this study resorts to variable importance and attribution measures (Section 2.4.1) to localise the relationships learnt by convolutional neural networks (CNN) between peripheral activity and individual brain regions. Rather than utilising *post hoc* explanation methods, which are not always faithful to the black-box models they attempt to explain [11], we can cast this setting into a GC inference problem and leverage GVAR in combination with CNN architectures. Expectedly, the applications to EEG data give rise to domain-specific research challenges and questions, e. g. the lack of perfect correspondence among electrode locations [196] across different study subjects and irregular time series sampling [197].

Another example summarised in Figure 3.6c is the analysis of wearables and self-reported symptoms data [194]. Xiao *et al.* [194] leverage sparse neural networks to identify groups of time series variables (modalities) most helpful in forecasting self-reported disease severity in chronic obstructive pulmonary disease (COPD) patients home-monitored long-term using multiple (mainly) wearable sensors [198]. Extraction of such predictive relationships and identification of the most important modalities (i) allow removing the burden from the doctors, caretakers, and patients by automating disease severity assessment and providing forecasts and (ii) facilitate cost reduction for the measurement setup in future studies and at deployment. The key challenge in applying GVAR in this context is the group structure among time series variables and the need for the joint selection of variable groups defined by modalities. While the GVAR's loss function includes the sparsity-inducing penalty, it can be, in addition, readily extended to the group [71] or adaptive structures [73].

To summarise, this section has outlined a few concrete applied problems from the biomedical domain (Figure 3.6), where time series structure learning and, specifically, our methods can be utilised. Given that the experiments described in the current chapter (Section 3.3) are limited to relatively simple benchmarks, tackling real-world scenarios, such as those outlined above, is one of the directions of future research. Naturally, differences and particularities in application domains and modelling problems pose additional challenges and requirements towards the models and inference frameworks, e. g. high dimensionality (Figure 3.6a), complex and varying measurement setups (Figure 3.6b), and multiple data sources (Figure 3.6c).

3.4.2 *Limitations*

The principal limitations of the GVAR model and inference framework are associated with the implicit assumptions generally related to Granger-causal inference [199]. Violations of these assumptions may lead to erroneous and misleading conclusions. For example, Granger-causal analysis assumes that the set of variables observed is *causally sufficient*, i. e. no unobserved confounders exist. Likewise, all the models in this chapter assume that the GC summary graph is fixed and does not evolve with time. Lastly, the inferred causal structures can be invalid when the methods are applied to under- or irregularly sampled time series without further precaution. Understanding, alleviating, and relaxing these restrictions comprise one fundamental research problem in causality.

Self-explaining neural networks at the basis of GVAR are closely related to the so-called *amortised explanation methods* [200]. Prior literature [200] has raised concerns regarding this class of methods. Specifically, it has been shown that such models tend to encode predictions into interpretations as part of the feature selection. We hypothesise that the GVAR is subject to similar limitations, and its coefficients may not be fully interpretable as they may encode forecast-related information.

From the computational perspective, the GVAR model (Equation 3.6) and forward-backward stability-based thresholding (Algorithm 1) present some tradeoffs compared to related neural-network-based GC inference methods. As mentioned above, we train $2K$ networks for the whole time series instead of utilising a separate neural network for each variable. While such an approach can considerably reduce the training time when the order of autoregressive relationships is smaller than the number of variables, it leads to higher memory complexity since the networks $\mathbf{\Phi}_{\boldsymbol{\theta}_k}$ have to output generalised coefficient matrices with $p \times p$ elements.

Another practical limitation generally applicable to all GC methods reliant on sparsity-inducing penalties is the challenges associated with hyperparameter tuning. Experimentally, we observed that the accuracy of the inferred causal structure was poorly correlated with the forecasting error on the held-out data. Thus, it may not be prudent to naïvely choose values for the parameters λ and γ (Equation 3.7) by minimising the validation-set error. Instead, it may be wiser to resort to more sophisticated model selection procedures, e. g. ones guided by stability [175]. Practical approaches to model selection in time series structure learning should constitute one of the research questions for future work.

Finally, as outlined in Section 3.4.1, a substantial limitation of the current experimental setup is the sole focus on synthetic benchmarking data with moderate dimensionality and simple dependency structures. A more diverse and well-rounded set of experiments is warranted to validate our methods, for instance, on higher-dimensional simulated gene regulatory networks [201] and in the real-world scenarios described in Section 3.4.1.

3.4.3 *Future Work*

The model and inference framework introduced in this chapter open many promising venues for future work. Empirically, as outlined in Sections 3.4.1–3.4.2, the current experimental setup should be extended to higher complexity and more realistic inference problems. Methodologically, many additional practical considerations can be incorporated into the method, for example, a principled treatment of variable groups [71] or even a learnable structure among the variables [73]. Since the GVAR produces *time-varying* generalised coefficient matrices, instead of assuming a fixed causal structure, it would be interesting to explore the inference of varying GC relationships [202], specifically in the context of highly nonstationary time series. The current implementation of GVAR resorts to utilising a linear link function and raw time series values as concepts (Definition 3.1.2). Using more specialised link functions, our framework can be applied to modelling point processes [203] and categorical time series [204]. Using and, possibly, learning concepts could provide a helpful inductive bias and reduce the model complexity associated with training a neural network for each lag. Beyond these model design adjustments, further performance improvements could be achieved by utilising more specialised architectures, for instance, CNNs applied in the frequency domain [193], [205] and recurrent networks in the spirit of the techniques by Tank *et al.* [95] and Khanna and Tan [96].

3.5 SUMMARY

This chapter tackled the design of interpretable models for handling temporal data and modelling autoregressive dependencies in the context of inferring nonlinear, multivariate Granger causality. Time series data are pertinent to many application domains, including biology and medicine, where many experiments and observations are conducted over long periods of time on a small set of selected measurement units. Understanding high-dimensional, nonlinear, and evolving cross-time dependencies among the covariates measured can shed light and provide a more systematic perspective on underlying processes and potentially causal relationships.

In particular, drawing on self-explaining neural networks and classical vector autoregression, we introduced a neural-network-based model, which extracts sparse and smoothly time-varying coefficients representing locally linear autoregressive relationships. Additionally, we developed a stability-based thresholding procedure that leverages time-reversed Granger causality to identify significant GC relationships based on the time-varying coefficients. We demonstrated the efficacy of this model and thresholding procedure in inferring GC relationships and the signs of interactions on several synthetic benchmarking datasets with known generative mechanisms and causal structures. Our experimental findings suggest that the proposed technique is a viable alternative to sparse neural networks, providing competitive structure inference results and being more interpretable and helpful for exploratory analysis.

4

PROTOTYPE-BASED EXPLANATIONS FOR DEEP SURVIVAL ANALYSIS

Next to the longitudinal observation treated in Chapter 3, a salient feature of healthcare datasets is the inclusion of outcome variables often encompassing time to adverse events, e. g. disease progression or recurrence and death. The techniques aimed at understanding associations between the observed covariates and survival time are broadly referred to as *survival analysis* [206], [207] and form a whole branch of biostatistics. The key distinguishing challenge of survival analysis is *censoring*: often, the outcome variable is observed only partially, for instance, for subjects who have withdrawn from the study.

Recent research efforts [208]–[213] have focused on developing neural-network-based approaches to survival analysis directly applicable to high-dimensional and unstructured inputs. A drawback of these powerful models is their opacity and lack of practitioner's insight into the nonlinear relationship between the covariates and time to event. In this chapter, we adopt a holistic perspective on deep survival analysis and introduce a fully probabilistic generative mixture model, allowing, in addition to classical survival regression, to conduct cluster analysis and utilise learnt clusters as prototype-based explanations (Section 2.4.4) elucidating the relationship between the covariates and survival time. The proposed model is *locally* interpretable and, owing to the previous advancements in stochastic gradient variational inference [105], readily scalable to high-dimensional and unstructured datasets.

In the remainder of this chapter, we provide the background behind survival regression and cluster analysis and generative and mixture modelling techniques. Subsequently, we introduce our method for interpretable neural-network-based survival analysis. After that, we describe the experimental setup and report empirical findings on several (semi-)synthetic and real-world problems, including an application to medical imaging data. We conclude with a detailed discussion of our contributions and findings. Most of this chapter is based on the contents and text of the published article "A Deep Variational Approach to Clustering Survival Data" [214].

4.1 BACKGROUND

To motivate the problem setting and justify model design choices, in this section, we introduce basic concepts relevant to the current chapter, such as survival analysis, generative modelling with variational autoencoders, and mixture regression models, providing a brief overview of the related work in these domains.

4.1.1 *Survival Analysis and Clustering*

The analysis of survival data is concerned with modelling the relationship between covariates, $\boldsymbol{x}$, and time to the event of interest, t, the latter being censored for some measurement units. Figure 4.1 shows an example of censored and uncensored survival times observed for a few high-grade glioma patients. In this case, t corresponds to the time to death elapsed from the patient's enrolment in the study.

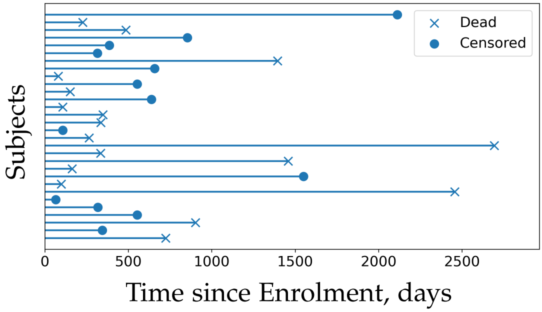


FIGURE 4.1: An example of survival times (in days) from the enrolment in the study for a subset of a high-grade glioma patient cohort (Section 4.3.1). Segments ending in "×" correspond to observing the event of interest, and "●" denotes censoring.

Survival data can be represented as tuples $\{(\boldsymbol{x}_i, \delta_i, t_i)\}_{i=1}^N$, where δ_i is the censoring indicator and t_i is the *observed* survival time. In this chapter, we limit our analysis and methods to so-called *right-censored* data [215]. Right censoring usually occurs when study subjects withdraw for external reasons. In particular, if the i-th data point was censored, $\delta_i = 0$ and the observed t_i corresponds to the censoring time, being a lower bound on the survival time. By contrast, in uncensored data points, $\delta_i = 1$ and t_i

equals the time to event. We assume that censoring is *noninformative*, i. e. independent of the subject's survival beyond censoring [206].

Survival analysis features a few quantities of interest. The *survival* function $S\left(t|\boldsymbol{x}\right)$ refers to the probability of the unit surviving beyond time t given its features $\boldsymbol{x}$ [216]. The *lifetime distribution* function is directly related to the survival function and is given by $F\left(t|\boldsymbol{x}\right) = 1 - S\left(t|\boldsymbol{x}\right)$. In turn, assuming F is differentiable, the *density* function is defined as $f\left(t|\boldsymbol{x}\right) = \frac{d}{dt}F\left(t|\boldsymbol{x}\right)$. Finally, the *hazard* function, corresponding to the instantaneous rate of event occurrence, is given by $h\left(t|\boldsymbol{x}\right) = \frac{f(t|\boldsymbol{x})}{S(t|\boldsymbol{x})}$.

Conventional statistical models for survival analysis impose strict assumptions on the form of the functions mentioned above. A classic example is the *proportional hazards* (PH) model introduced in the seminal work by Cox [216], which directly defines the hazard function as follows:

$$h\left(t|\boldsymbol{x}\right) = h_0\left(t\right)\exp\left\{\boldsymbol{\beta}^\top\boldsymbol{x}\right\}, \tag{4.1}$$

where h_0 is the baseline hazard function corresponding to $\boldsymbol{x} = \boldsymbol{0}$, and $\boldsymbol{\beta}$ is the coefficient vector. In Equation 4.1, the term $\exp\left\{\boldsymbol{\beta}^\top\boldsymbol{x}\right\}$ is the *proportionate* increase or decrease in risk for the data point with covariates $\boldsymbol{x}$.

Arguably, an even more direct approach is adopted by the accelerated failure time (AFT) models [217], where, instead of having a multiplicative effect on the hazard function (Equation 4.1), $\boldsymbol{x}$ acts multiplicatively on the time:

$$t = \exp\left\{\boldsymbol{\beta}^\top\boldsymbol{x}\right\}t_0, \tag{4.2}$$

where t_0 denotes an exponentiated error term. Interestingly, the two models defined above (Equations 4.1–4.2) can be reconciled by assuming that the survival times follow a Weibull distribution [206] with the density function given by:

$$f(t;\lambda,k) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}\exp\left\{-\left(\frac{t}{\lambda}\right)^k\right\}, \tag{4.3}$$

where $\lambda, k > 0$ are the scale and shape parameters of the distribution. In practice, the dependency on the covariates is modelled by representing these distribution parameters as simple functions of $\boldsymbol{x}$.

The parameters of the survival models, such as AFT and Cox PH, are conventionally optimised by maximising the likelihood or partial likelihood function [218], which, assuming given right-censored and IID data, can be written as

$$\prod_{i=1}^{N} f\left(t_i|\boldsymbol{x}_i\right)^{\delta_i} S\left(t_i|\boldsymbol{x}_i\right)^{1-\delta_i} = \prod_{i=1}^{N} h\left(t_i|\boldsymbol{x}_i\right)^{\delta_i} S\left(t_i|\boldsymbol{x}_i\right). \tag{4.4}$$

Observe that if the i-th data point was not censored, its contribution to the likelihood above amounts to the factor of $f\left(t_i|\boldsymbol{x}_i\right)$. By contrast, a censored data point contributes a factor of $S\left(t_i|\boldsymbol{x}_i\right)$, which, by definition, corresponds to the probability of surviving beyond the observed censoring time.

Cox PH and AFT models assume a simple log-linear relationship between the covariates and the hazard function or survival time (Equations 4.1–4.2). Machine learning literature has explored alternative methods to alleviate this and other limitations. For example, building on the success of random forests [4], Ishwaran *et al.* [219] apply tree-based ensembles to survival regression. Instead of the conventional linear predictor in Equation 4.1, Faraggi and Simon [208] utilise neural networks to map subjects' features to risk. A similar approach with modern network architectures and a focus on treatment recommendation is investigated by Katzman *et al.* [210]. Kvamme, Borgan, and Scheel [212] extend this framework beyond the PH assumption. By contrast, Ranganath *et al.* [209] adopt a Bayesian generative perspective using deep latent variable models [220] to relate the covariates and survival time. Yet another family of methods treats the time to event as a discrete ordinal variable and tackles the more advanced setting of competing risks [211]. A research direction closely related to the current chapter is mixtures of regression models and methods for the cluster analysis of survival data, briefly overviewed in the remainder of this section.

Cluster analysis of survival data is an under-explored research problem with applications to exploratory data analysis and patient subgroup identification [221]. In contrast to the conventional survival analysis, where the quantities of interest mainly relate the covariates to the survival time, in clustering, we additionally consider an unobserved (latent) cluster assignment variable, denoted by c throughout this chapter. In this scenario, the challenge is twofold: (i) inferring the unobserved cluster assignment and (ii) modelling the survival time distribution given the covariates and inferred cluster label. Thus, for example, for the i-th data point, the survival function is given by $S\left(t_i|\boldsymbol{x}_i, c_i\right)$, where time is conditioned on the cluster assignment and the covariates.

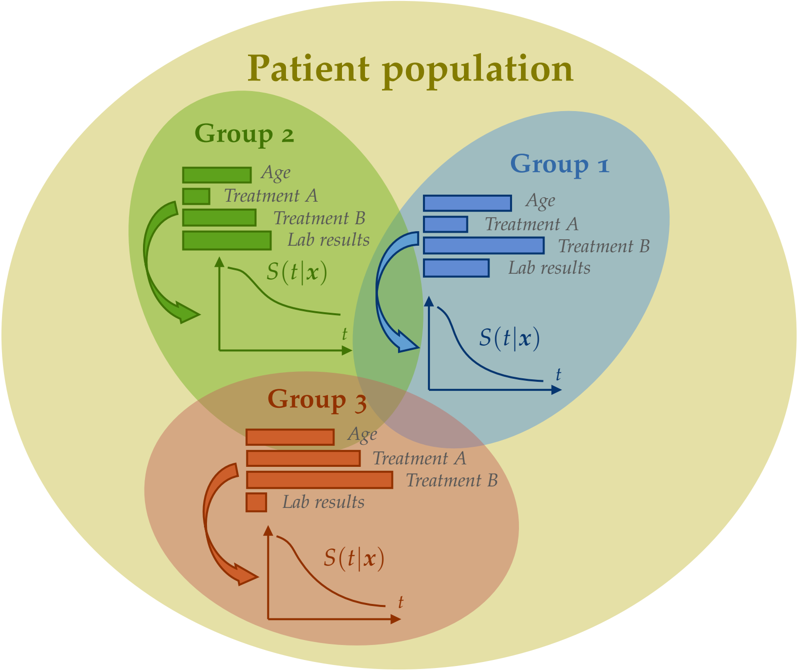


FIGURE 4.2: A schematic summary of the survival cluster analysis. The patient population consists of three groups with disparate survival functions and varying relationships between the covariates and survival. The task is to (i) infer cluster assignments and (ii) model the time to event given the covariates and cluster label.

Figure 4.2 shows a schematic summary of the survival cluster analysis. Herein, the general **patient population** comprises three groups, or clusters, characterised by varying survival functions and relationships between the covariates and time to event. For instance, in **group 3**, treatment *A* has a considerably stronger effect on survival than in **group 2**. In addition to the survival functions and covariate effects, the groups may present with different feature distributions similar to the conventional unsupervised clustering setting.

The survival cluster analysis problem outlined above is especially relevant from the perspective of precision medicine [222], which seeks a more personalised approach to patient management. Opposed to classical unsupervised clustering, which is not guaranteed to discover groups related to differences in survival time distributions and covariate effects [223], a semisupervised learning approach would allow to jointly consider survival times and features and discover structures such as the one summarised in Figure 4.2. In general, identifying such patient subpopulations from healthcare data can facilitate a deeper understanding of diseases and the development of novel management strategies [224].

Several lines of related work tackle the problem of clustering and group-based survival analysis. Bair and Tibshirani [223] introduce a semisupervised clustering technique comprising preselection of predictors correlated with survival time using a Cox PH model and k-means clustering performed on the chosen feature subset. By contrast, Ahlqvist *et al.* [221] apply Cox PH regression to the groups discovered by k-means and hierarchical clustering in a cohort of diabetic patients. Another line of research focuses on differentiating between long- and short-term survivors using mixtures of regression models and clustering methods [225], [226]. Xia *et al.* [227] adopt a multitask approach to outcome-driven clustering of acute coronary syndrome patients.

Beyond the efforts mentioned above, a few works are very closely related to the scope of the current chapter. Chapfuwa *et al.* [228] introduce survival cluster analysis (SCA), which leverages a truncated Dirichlet process and time-to-event prediction on neural network representations. Similarly, deep survival machines (DSM) by Nagpal, Li, and Dubrawski [213] fit a finite mixture of Weibull models on the embeddings produced by an encoding neural network. A similar class of models is deep Cox mixtures (DCM) [229], which also regularise the network's representations by imposing a prior distribution and utilise a mixture of Cox regression instead of the Weibull models. Last but not least, Liverani *et al.* [230] introduce a fully generative clustering method for collinear survival data building on the Bayesian profile regression (PR) [231]. This method equips a Dirichlet process with a mixture of cluster-specific Weibull models.

4.1.2 *Variational Autoencoders*

The model for explainable nonlinear survival analysis introduced in this chapter (Section 4.2) is a deep *latent variable model* (LVM) [232], [233] mainly based on variational autoencoders (VAE) [105] and their extensions to unsupervised clustering [234], [235]. Below, we provide a basic introduction to this class of deep generative models.

In the spirit of representation learning [104], LVMs assume that observed variables $\boldsymbol{x}$ can be directly explained by *latent* (unobserved) variables $\boldsymbol{z}$,

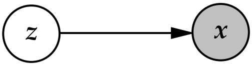


FIGURE 4.3: Directed acyclic graph representing generative assumptions of a latent variable model. Shaded nodes correspond to observed variables, and unshaded nodes denote unobserved (latent) variables.

where the latter typically correspond to high-level generative factors. The generative process assumed by LVMs is summarised as a graphical model in Figure 4.3. More formally, first, latent variables $\boldsymbol{z}$ are sampled from a *prior distribution* $p\left(\boldsymbol{z}\right)$. Subsequently, the observed covariates $\boldsymbol{x}$ are sampled from a conditional distribution $p\left(\boldsymbol{x}|\boldsymbol{z}\right)$. Thus, the likelihood of the observed data is given by

$$p\left(\boldsymbol{x}\right)=\int p\left(\boldsymbol{x}|\boldsymbol{z}\right)p\left(\boldsymbol{z}\right)d\boldsymbol{z}. \tag{4.5}$$

Classic examples of latent variable modelling include probabilistic principal component [232], factor, and latent class analyses [236].

As mentioned, a variational autoencoder [105] is a *deep* latent variable model in that the conditional distribution $p\left(\boldsymbol{x}|\boldsymbol{z}\right)$ is parameterised by a (deep) neural network [237] and, therefore, is denoted by $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)$. The underlying network is often referred to as a *decoder* or *generative network*. In this setting, the integral in Equation 4.5 is intractable, and so is the *posterior distribution* $p_{\boldsymbol{\theta}}\left(\boldsymbol{z}|\boldsymbol{x}\right)=p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)p_{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)/p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$. Hence, VAEs resort to *amortised variational inference* [238], where the true posterior is approximated by the distribution $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)$ with *variational parameters* $\boldsymbol{\phi}$. To avoid per-data-point optimisation, similar to the generative model, the variational posterior is parameterised by a neural network called an *encoder* or *recognition network*.

In practice, VAE model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are optimised by maximising a lower bound on the log-likelihood of the observed data (Equation 4.5), namely:

$$\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\geq\mathbb{E}_{q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)}\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)-D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)\|p_{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\right), \tag{4.6}$$

where D_{KL} denotes the Kullback–Leibler (KL) divergence. Intuitively, the term $\mathbb{E}_{q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)}\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)$ corresponds to the reconstruction loss, whereas the KL divergence is a regularisation term. The RHS in Equation 4.6 is the *evidence lower bound* (ELBO); its step-by-step derivation is provided in Appendix B.1.

The objective above is typically simplified by assuming the standard multivariate normal distribution as the prior, denoted by $p\left(\boldsymbol{z}\right)$, and a Gaussian distribution with a diagonal covariance matrix for $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)$. The ELBO allows optimising all parameters jointly using gradient methods. In particular, differentiable sampling is facilitated by the so-called *reparameterisation trick* [105], [239].

Salient advantages of the framework introduced by Kingma and Welling [105] are that it can be readily adapted beyond the vanilla representation learning scenario and that practitioner's generative assumptions can be explicitly and interpretably specified as, for instance, graphical models [240] and incorporated into the learning. As a result, many extended VAE variants have been considered in the literature, e. g. works by Kingma *et al.* [241], Sohn, Lee, and Yan [242], and Wu and Goodman [243] are just a few examples. Particularly relevant to the scope of this chapter are the methods for generative unsupervised clustering introduced by Dilokthanakul *et al.* [234] and Jiang *et al.* [235].

Herein, we will briefly introduce the *variational deep embedding* (VaDE) model [235], which helps understand the design choices behind the method in Section 4.2. VaDE combines VAEs with Gaussian mixture models (Section 4.1.3) by introducing an additional latent cluster assignment variable $c \in \{1, \ldots, K\}$, thus facilitating end-to-end generative unsupervised clustering. In VaDE, the posterior distribution is given by $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}, c|\boldsymbol{x}\right)$, and the ELBO may be written as

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z},c|\boldsymbol{x})} \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right) - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z}, c|\boldsymbol{x}\right) \| p\left(\boldsymbol{z}, c\right)\right), \tag{4.7}$$

where $p\left(\boldsymbol{z}, c\right)$ is the Gaussian mixture prior. Note that, similar to the ELBO of VAE (Equation 4.6), the first term in Equation 4.7 corresponds to the reconstruction loss, whereas the second term regularises the posterior distribution, assuming the mixture of Gaussians prior. A detailed discussion of the generative process and other assumptions made by VaDE is omitted for brevity; however, we provide a commentary on some of these aspects in the following sections.

4.1.3 *Mixture Models*

Yet another class of methods employing latent variables (Figure 4.3) are mixture models [244]. Generally, mixture models are useful for clustering and nonlinear regression, building on the assumption that observations originate from several distinct subpopulations. In mixture models, the latent variables are discrete and correspond to the cluster (group) assignments. The current literature contains many variants of these models extending to (semi-)supervised learning scenarios, for example, *mixtures of experts* [245], [246] and *mixtures of regression models* [247]. As mentioned in Section 4.1.1, a few related works also explore mixture modelling approaches to survival analysis [213], [228]–[230].

In the unsupervised setting, a classic example is the *Gaussian mixture model* (GMM) [248], leveraged for clustering by the VaDE discussed in Section 4.1.2. As in other LVMs (Equation 4.5), for GMMs, the likelihood of the observed data is given by $\sum_{c=1}^{K} p\left(\boldsymbol{x}|c\right) p\left(c\right)$. Herein, the prior is a categorical distribution $p\left(c\right) = \pi_c$, where, for $c \in \{1, \ldots, K\}$, $\pi_c \geq 0$ and $\sum_{c=1}^{K} \pi_c = 1$. Assuming $\boldsymbol{x} \in \mathbb{R}^p$, the conditional distribution is multivariate normal:

$$p\left(\boldsymbol{x}|c\right) = \frac{1}{\sqrt{(2\pi)^p \left|\boldsymbol{\Sigma}_c\right|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_c\right)^\top \boldsymbol{\Sigma}_c^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_c\right)\right\}, \qquad (4.8)$$

where $\boldsymbol{\mu}_c \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_c \in \mathbb{R}^{p \times p}$ are the cluster-specific mean vector and covariance matrix; additional inductive biases may be introduced into the generative model by imposing assumptions on the structure of the cluster-specific covariance matrices. Like for other LVMs, the inference problem amounts to estimating the posterior distribution $p\left(c|\boldsymbol{x}\right)$ and is typically solved using the expectation–maximisation (EM) algorithm [249].

4.2 VARIATIONAL DEEP SURVIVAL CLUSTERING

The chief contribution of this chapter is a novel probabilistic approach to deep survival and cluster analysis (Figure 4.2). The model introduced in this section allows for the interpretation of nonlinear relationships between the covariates and survival time in terms of cluster assignments and cluster-specific parameters, which effectively comprise prototype-based explanations. Our *variational deep survival clustering* (VaDeSC) model builds on the techniques for unsupervised clustering in the deep variational setting [235] and approaches for deep survival analysis and mixture modelling [209], [213], [230] by augmenting a generative clustering model with cluster-specific survival functions. This framework can capture heterogeneities in the distributions of covariates *and* survival times, as well as *varying relationships* between the two. Figure 4.4 provides a schematic overview of the VaDeSC model. Note that, for legibility, the figure depicts a concrete example with two clusters, although the model is generally applicable to any given number of components. In the following sections, we describe the model's generative assumptions and structure, following the notation and terminology introduced in Section 4.1.

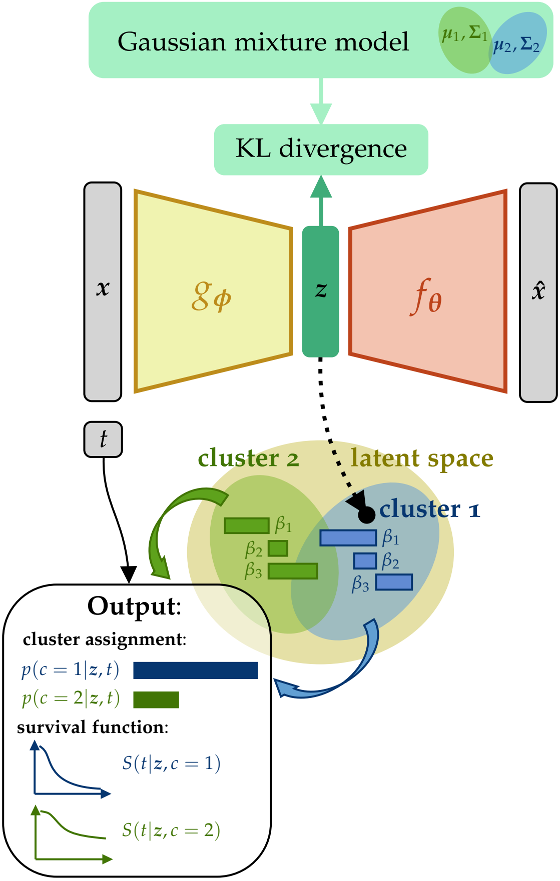


FIGURE 4.4: A schematic overview of the variational deep survival clustering model. An encoder neural network (g_{ϕ}) maps covariate vectors ($\boldsymbol{x}$) to low-dimensional latent representations ($\boldsymbol{z}$). The representations are regularised by a mixture of Gaussians prior. Each observation is assigned to a cluster (c) conditional on the representation and observed survival time (t). The latent space is equipped with Weibull models, which relate representations to the scale parameter of the survival time distribution via cluster-specific coefficients ($\boldsymbol{\beta}$). Finally, the decoder (f_{θ}) reconstructs representations in the feature space ($\hat{\boldsymbol{x}}$).

4.2.1 *Generative Model*

The generative model is based on the informal definition of the survival cluster analysis problem outlined in Section 4.1 and motivated by prototype-based explanations (Section 2.4.4), specifically, the Bayesian case model [138]. Thus, the variational deep survival clustering with K components assumes the following generative process per data point:

1. Sample a cluster assignment $c \sim \text{Categorical}\left(\boldsymbol{\pi}\right)$, where $\boldsymbol{\pi} = \begin{pmatrix} \pi_1 & \cdots & \pi_K \end{pmatrix}^\top$, $\pi_j > 0$, for $1 \leq j \leq K$, are prior cluster probabilities s.t. $\sum_{j=1}^{K} \pi_j = 1$. We will use $p(c)$ as a shorthand for this categorical prior distribution.

2. Sample a latent representation $\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)$, where $\boldsymbol{\mu} = \left\{\boldsymbol{\mu}_j\right\}_{j=1}^{K}$ and $\boldsymbol{\Sigma} = \left\{\boldsymbol{\Sigma}_j\right\}_{j=1}^{K}$ are cluster-specific mean vectors and covariance matrices. Let $p\left(\boldsymbol{z}|c\right)$ denote the distribution over latent variables given the cluster assignment.

3. Sample a feature vector $\boldsymbol{x} \sim p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)$, where the distribution $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)$ is parameterised by a decoder $f_{\boldsymbol{\theta}}$. The conditional distribution needs to be chosen depending on the data type. For example, for continuous covariates, we assume that $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \text{diag}\left(\boldsymbol{\sigma}_{\boldsymbol{\theta}}\right)\right)$, where $\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\sigma}_{\boldsymbol{\theta}}\right) = f_{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)$. By contrast, for binary covariates, we assume $\boldsymbol{x} \sim \text{Bernoulli}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\right)$. Other design choices may be appropriate for alternative data types, e. g. for count or categorical features.

4. Sample a survival time $t \sim p\left(t|\boldsymbol{z}, c\right)$. Observe that the distribution is cluster-specific; details are provided in Section 4.2.2. Similar to the conventional survival analysis (Section 4.1.1), the observed time is censored depending on the indicator δ. We assume that censoring is *noninformative*.

The generative process above resembles that of the VaDe by Jiang *et al.* [235] in combining variational autoencoder and Gaussian mixture models (Sections 4.1.2–4.1.3). Figure 4.5 summarises data-generating assumptions as a graphical model using plate notation. Note that some survival distribution parameters shown in the graph are introduced in Section 4.2.2. Importantly, these assumptions induce a helpful factorisation of the likelihood described in Section 4.2.3 and allow utilising clusters and mean vectors as prototypes in explaining the complex relationship between $\boldsymbol{x}$ and t.

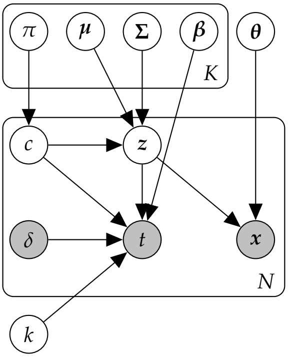


FIGURE 4.5: Directed acyclic graph representing generative assumptions of the variational deep survival clustering model. The plates correspond to cluster-specific parameters (K) and per-data-point variables (N). Shaded nodes correspond to observed variables, and unshaded nodes denote latent variables and parameters.

4.2.2 *Survival Model*

The generative process above (Figure 4.5) assumes that the observed survival time depends directly on the censoring indicator, cluster assignment, and latent representation. Similar to several closely related approaches [209], [213], [230], we assume that conditional on the variables mentioned, the uncensored survival time is sampled from the Weibull distribution (Equation 4.3) with the shape parameter $k > 0$ and the scale parameter determined by the latent variables z and cluster assignment c. Including our assumptions about the censoring mechanism (cf. Equation 4.4), the conditional distribution of the survival time is given by

$$\begin{aligned} p\left(t|z,c\right) &= f\left(t;\lambda\left(z,c\right),k\right)^{\delta} S\left(t|z,c\right)^{1-\delta} \\ &= \left[\frac{k}{\lambda\left(z,c\right)}\left(\frac{t}{\lambda\left(z,c\right)}\right)^{k-1}\exp\left\{-\left(\frac{t}{\lambda\left(z,c\right)}\right)\right\}\right]^{\delta} \\ &\qquad \cdot\left[\exp\left\{-\left(\frac{t}{\lambda\left(z,c\right)}\right)^{k}\right\}\right]^{1-\delta}, \end{aligned} \tag{4.9}$$

where the scale parameter is defined as follows:

$$\lambda\left(z,c\right) = \mathrm{softplus}\left(\boldsymbol{\beta}_c^{\top} z\right) = \log\left(1+\exp\left\{\boldsymbol{\beta}_c^{\top} z\right\}\right). \tag{4.10}$$

In Equation 4.10, $\boldsymbol{\beta}=\left\{\boldsymbol{\beta}_j\right\}_{j=1}^{K}$ denote cluster-specific coefficient vectors, which relate the latent space to the scale of the Weibull distribution via the softplus function, a smooth version of the rectified linear unit (ReLU). Note that, in the first line of Equation 4.9, $S\left(t|\boldsymbol{z},c\right)=\int_{\tau=t}^{\infty}f\left(\tau,\lambda\left(\boldsymbol{z},c\right),k\right)d\tau$. Under the specified distribution, the expected uncensored survival time is $\lambda\left(\boldsymbol{z},c\right)\cdot\Gamma\left(1+1/k\right)$, and the median is $\lambda\left(\boldsymbol{z},c\right)\cdot\left(\ln 2\right)^{1/k}$, where Γ is the gamma function.

One design choice reflected by Equation 4.9 is that the scale parameter $\lambda\left(\boldsymbol{z},c\right)$ is cluster- and instance-specific, whereas the shape parameter k is global. More flexible formulations are viable; for example, in the mixture model by Liverani *et al.* [230], shape parameters are learnable and cluster-specific. The assumption of the Weibull distribution is yet another design decision and is justified by this distribution's unique property of being both proportional and accelerated (Section 4.1.1) [206], [250]. In this specific, other survival time modelling approaches are possible, e. g. mixtures of semiparametric models proposed by Nagpal *et al.* [229].

4.2.3 *Evidence Lower Bound and Optimisation*

Following generative and model assumptions from Sections 4.2.1–4.2.2, the likelihood of the observed data can be written as

$$\begin{aligned}p\left(\boldsymbol{x},t\right)&=\int\sum_{c=1}^{K}p\left(\boldsymbol{x}|t,\boldsymbol{z},c\right)p\left(t,\boldsymbol{z},c\right)d\boldsymbol{z}\\&=\int\sum_{c=1}^{K}p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right)p\left(t|\boldsymbol{z},c\right)p\left(\boldsymbol{z}|c\right)p\left(c\right)d\boldsymbol{z},\end{aligned}\tag{4.11}$$

where the factorisation in the second line follows from the conditional independences implied by the graph in Figure 4.5. Observe that the likelihood function above corresponds to a single observation, and a product of terms needs to be considered for a dataset of IID points. Note that Equations 4.9 and 4.11 treat censoring indicators as fixed inputs. Finally, recall that $p\left(t|\boldsymbol{z},c\right)$ depends on the cluster-specific parameters $\boldsymbol{\beta}$, $p\left(\boldsymbol{z}|c\right)$ depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and $p\left(c\right)$ is defined based on $\boldsymbol{\pi}$. We have omitted these dependencies in our notation for legibility.

Following the approach discussed in Section 4.1.2, we resort to amortised variational inference to approximate the intractable posterior distribution $p_{\boldsymbol{\theta}}\left(\boldsymbol{z},c|\boldsymbol{x},t\right)$ with $q_{\boldsymbol{\phi}}\left(\boldsymbol{z},c|\boldsymbol{x},t\right)$ parameterised by an encoder neural network

g_{ϕ}. As in Appendix B.1, we derive a lower bound on the log-likelihood of the observed data:

$$\log p\left(x,t\right)=\log\int\sum_{c=1}^{K}p_{\theta}\left(x|z\right)p\left(t|z,c\right)p\left(z|c\right)p\left(c\right)\frac{q_{\phi}\left(z,c|x,t\right)}{q_{\phi}\left(z,c|x,t\right)}dz \tag{4.12}$$

$$\geq\mathbb{E}_{q_{\phi}(z,c|x,t)}\log\frac{p_{\theta}\left(x|z\right)p\left(t|z,c\right)p\left(z|c\right)p\left(c\right)}{q_{\phi}\left(z,c|x,t\right)}, \tag{4.13}$$

where Equation 4.13 is implied by the Jensen's inequality. Similar to the VaDE model by Jiang *et al.* [235], we assume the following factorisation of the variational distribution:

$$q_{\phi}\left(z,c|x,t\right):=q_{\phi}\left(z|x\right)q\left(c|z,t\right), \tag{4.14}$$

where, intuitively, $q_{\phi}\left(z|x\right)$ corresponds to the encoder and $q\left(c|z,t\right)$ is the soft cluster assignment. Equation 4.14 is somewhat similar to the factorisation proposed by Jiang *et al.* [235] based on the mean-field assumption. Although the latter does not fully translate to the specific factorisation above, we observe that the performance of the VaDeSC in the experiments is encouraging (Section 4.4). Moreover, under this factorisation, the encoding and cluster assignment are separated, and representations are not conditioned on the survival time, which is typically missing at inference.

Another trick we borrow from the VaDE [235] is to replace the variational approximation $q\left(c|z,t\right)$ from Equation 4.14 with the distribution $p\left(c|z,t\right)$:

$$\begin{aligned}q\left(c|z,t\right):=p\left(c|z,t\right)&=\frac{p\left(z,t|c\right)p\left(c\right)}{\sum_{c'=1}^{K}p\left(z,t|c'\right)p\left(c'\right)}\\&=\frac{p\left(t|z,c\right)p\left(z|c\right)p\left(c\right)}{\sum_{c'=1}^{K}p\left(t|z,c'\right)p\left(z|c'\right)p\left(c'\right)}.\end{aligned} \tag{4.15}$$

Alternatively, $q\left(c|z,t\right)$ may be parameterised by a neural network classifier with learnable parameters. Nevertheless, we resort to Equation 4.15 to reduce computational costs and mitigate potential training instability and overfitting.

Finally, as for the conventional VAEs [105], we assume a multivariate normal distribution with a diagonal covariance matrix for $q_{\phi}\left(z|x\right)$. Thus, the output of the encoder is given by $\left(\mu_{\phi},\sigma_{\phi}\right)=g_{\phi}\left(x\right)$, and the resulting variational posterior corresponds to $\mathcal{N}\left(\mu_{\phi},\mathrm{diag}\left(\sigma_{\phi}\right)\right)$.

Given all the assumptions explained above, the ELBO from Equation 4.13 can be rewritten in the following form:

$$\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\Big[\underbrace{\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})}_{❶}+\underbrace{\log p(t|\boldsymbol{z},c)}_{❷}+\underbrace{\log p(\boldsymbol{z}|c)}_{❸}+\underbrace{\log p(c)}_{❹}\\
&-\underbrace{\log q_{\boldsymbol{\phi}}(\boldsymbol{z},c|\boldsymbol{x},t)}_{❺}\Big]=\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})+\log p(t|\boldsymbol{z},c)\Big] \qquad (4.16)\\
&\qquad\qquad\qquad\qquad\qquad -D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z},c|\boldsymbol{x},t)\,\|\,p(\boldsymbol{z},c)\right).
\end{aligned}$$

Note the resemblance between Equations 4.16 and 4.7. Similar to the VaDE, the ELBO of the VaDeSC model includes a reconstruction loss and the Kullback–Leibler divergence, regularising the variational distribution. Below, we comment on each of the terms of the lower bound marked by black circles in Equation 4.16.

❶ The first summand is the conventional *reconstruction* term. Similar to the rest of the ELBO terms, it can be approximated using the stochastic gradient variational Bayes (SGVB) estimator [105]. Assumptions on the distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ comprise one of the design choices in our method and are determined by the type of covariates (Section 4.2.1). For example, assuming $\boldsymbol{x}$ consists of p *binary* features, the reconstruction term can be specified and approximated as

$$\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})=\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\\
&\quad\approx\frac{1}{L}\sum_{l=1}^{L}\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}^{(l)}\right)=\frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{p}\Big\{x_j\log\mu_{\boldsymbol{\theta},j}^{(l)} \qquad (4.17)\\
&\qquad\qquad\qquad\qquad\qquad +(1-x_j)\log\left(1-\mu_{\boldsymbol{\theta},j}^{(l)}\right)\Big\},
\end{aligned}$$

where $\boldsymbol{z}^{(l)}$, for $1\leq l\leq L$, are sampled from the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(l)}=f_{\boldsymbol{\theta}}\left(\boldsymbol{z}^{(l)}\right)$ is the output of the decoder.

❷ The second term introduces *supervision* by considering the conditional distribution of the survival time. As explained in Section 4.2.2, we assume a mixture of cluster-specific Weibull survival models. The SGVB estimator for this term is given by

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\log p(t|\boldsymbol{z},c)\approx\sum_{l=1}^{L}\sum_{c'=1}^{K}p\left(c'|\boldsymbol{z}^{(l)},t\right)\log p\left(t|\boldsymbol{z}^{(l)},c'\right), \qquad (4.18)$$

where $p\left(t|\boldsymbol{z}^{(l)},c'\right)$ is defined according to Equation 4.9 and $p\left(c'|\boldsymbol{z}^{(l)},t\right)$ is computed as in Equation 4.15.

③ The next term is related to the *clustering* loss. The distribution of the representations conditional on the cluster assignments is given by a mixture of Gaussians:

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\log p\left(\boldsymbol{z}|c\right)&=\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\sum_{c'=1}^{K}p\left(c'|\boldsymbol{z},t\right)\log p\left(\boldsymbol{z}|c'\right)\\&\approx\frac{1}{L}\sum_{l=1}^{L}\sum_{c'=1}^{K}p\left(c'|\boldsymbol{z}^{(l)},t\right)\log p\left(\boldsymbol{z}^{(l)}|c'\right),\end{aligned}\tag{4.19}$$

where $p\left(\boldsymbol{z}^{(l)}|c'\right)$ is the density of the multivariate normal distribution (Equation 4.8) with the mean $\boldsymbol{\mu}_{c'}$ and covariance matrix $\boldsymbol{\Sigma}_{c'}$.

④ The fourth term includes the *prior* distribution on cluster assignments and is approximated as

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\log p\left(c\right)\approx\frac{1}{L}\sum_{l=1}^{L}\sum_{c'=1}^{K}p\left(c'|\boldsymbol{z}^{(l)},t\right)\log p\left(c'\right).\tag{4.20}$$

Recall that $p(c)$ is a categorical distribution with probabilities $\boldsymbol{\pi}$ (Section 4.2.1). In our experiments (Section 4.3), we treat these probabilities as fixed parameters by utilising the uniform distribution as the prior, i. e. we set $\pi_j = 1/K$ for $1 \leq j \leq K$.

⑤ Lastly, the fifth term of the ELBO corresponds to the *entropy* of the variational distribution:

$$\begin{aligned}-\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\log q_{\boldsymbol{\phi}}\left(\boldsymbol{z},c|\boldsymbol{x},t\right)&=-\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\big[\log q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)\\+\log p\left(c|\boldsymbol{z},t\right)\big]&=-\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\log q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)\\&\quad-\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})p(c|\boldsymbol{z},t)}\log p\left(c|\boldsymbol{z},t\right),\end{aligned}\tag{4.21}$$

where the first summand is the entropy of $q_{\boldsymbol{\phi}}\left(\boldsymbol{z}|\boldsymbol{x}\right)$, a multivariate normal distribution with a diagonal covariance matrix. Assuming the latent space with J dimensions, Equation 4.21 can be approximated by

$$\frac{J}{2}\log\left(2\pi e\right)+\sum_{j=1}^{J}\log\sigma_{\boldsymbol{\phi},j}^{2}-\frac{1}{L}\sum_{l=1}^{L}\sum_{c'=1}^{K}p\left(c'|\boldsymbol{z}^{(l)},t\right)\log p\left(c'|\boldsymbol{z}^{(l)},t\right).\tag{4.22}$$

To train the VaDeSC model, we maximise the objective in Equations 4.13 and 4.16 using SGVB estimators given by Equations 4.17–4.21. Specifically, the problem amounts to

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z},c|\boldsymbol{x},t)} \log \frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{x}|\boldsymbol{z}\right) p\left(t|\boldsymbol{z},c\right) p\left(\boldsymbol{z}|c\right) p\left(c\right)}{q_{\boldsymbol{\phi}}\left(\boldsymbol{z},c|\boldsymbol{x},t\right)}, \tag{4.23}$$

where $\boldsymbol{\mu} = \left\{\boldsymbol{\mu}_j\right\}_{j=1}^{K}$, $\boldsymbol{\Sigma} = \left\{\boldsymbol{\Sigma}_j\right\}_{j=1}^{K}$, and $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_j\right\}_{j=1}^{K}$ are cluster-specific learnable parameters, as explained in Sections 4.2.1–4.2.2. Recall that the prior distribution $p\left(c\right)$ is fixed. In summary, the model's parameters can be optimised jointly using practical approaches for gradient-based learning, such as minibatch stochastic gradient descent.

4.2.4 *Further Remarks*

Cluster assignment c can help *interpret* the relationship between the covariates $\boldsymbol{x}$ and (observed or predicted) survival time t. In particular, instances assigned to the cluster of interest and other clusters can be provided as case-based explanations. Moreover, we can visualise *generated* instances belonging to different clusters by sampling the mixture of Gaussians prior and decoding the sampled representations (Section 4.2.1). Last but not least, similarly, we can apply the decoder to the mean vector of the assigned cluster $\boldsymbol{\mu}_c$ and return the decoded features as a prototype-based explanation.

Given a feature vector $\boldsymbol{x}$ and survival time t, the cluster assignment is determined by the distribution $p\left(c|\boldsymbol{z},t\right)$ computed according to Equation 4.15, which, in turn, depends on the survival time distribution $p\left(t|\boldsymbol{z},c\right)$. However, at test time, t may not be observed and we resort to using the following distribution for cluster assignment:

$$p\left(c|\boldsymbol{z}\right) = \frac{p\left(\boldsymbol{z}|c\right) p\left(c\right)}{\sum_{c'=1}^{K} p\left(\boldsymbol{z}|c'\right) p\left(c'\right)}. \tag{4.24}$$

Observe that Equation 4.24 follows from the Bayes' rule and none of the terms above depends on the unobserved survival time.

4.3 EXPERIMENTAL SETUP

This section provides an overview of the experimental setup applied to produce empirical findings presented in Section 4.4. The goal of our experiments is severalfold: (i) to validate the VaDeSC model on benchmarking datasets with *known* clustering structure, (ii) to evaluate the predictive performance of our model in survival analysis tasks of varying complexity, (iii) to illustrate how the high-dimensional and nonlinear relationships learnt by the model can be interpreted using the discovered clusters, and (iv) to demonstrate the utility of the method on *real-world* medical data. Below, we describe the benchmarking datasets, baseline methods to which we compare the VaDeSC model, and ablations and introduce some specialised evaluation metrics.

4.3.1 *Datasets*

The experiments are performed on several benchmarks, comprising synthetic and real-world survival data. The datasets represent scenarios with different numbers of data points, explanatory variables, frequency of censored observations, data types, and clustering structures. Below, we describe the benchmarks, with a summary displayed in Table 4.1.

Dataset	N	p	**% censored**	**Data type**	K	**Balanced?**
Synthetic	60000	1000	30	Tabular	3	✓
survMNIST	70000	28×28	52	Image	5	✗
SUPPORT	9105	59	32	Tabular	—	—
FLChain	6524	7	70	Tabular	—	—
HGG	453	147	25	Tabular	—	—
Hemodialysis	1493	57	91	Tabular	—	—
NSCLC	961	64×64	33	Image	—	—

TABLE 4.1: A summary of benchmarking datasets included in the experiments. The abbreviations are introduced in the remainder of this section. For every dataset, we report the total number of data points (N), feature dimensionality (p), percentage of censored survival times, number of ground-truth clusters (K), and whether the cluster sizes are balanced. Note that, in most cases, the clustering structure is unknown.

To investigate a controlled and simple benchmarking problem, we include experiments on nonlinear **synthetic** data simulated using a procedure similar to the generative process assumed by the VaDeSC model (Section 4.2.1). In particular, low-dimensional representations are sampled from a mixture of Gaussians and mapped to high-dimensional feature vectors and survival times using randomly initialised MLPs with nonlinear activation functions. Appendix B.2 contains a detailed step-by-step description of the generation procedure and relevant parameters. For this dataset, we hold out 18000 data points as the test set and consider five independent simulations.

Another synthetic dataset with a known clustering structure is the survival MNIST (**survMNIST**), adapted from the benchmark introduced by Pölsterl [251], which is based on the famous Modified National Institute of Standards and Technology (MNIST) handwritten digit database [2]. In this toy problem, low-resolution images of handwritten digits serve as explanatory variables. Every digit is assigned to a cluster and is accompanied by a synthetic survival time (Appendix B.2). In our experiments, we utilise the train-test split from the original MNIST dataset with 10000 data points in the test set and consider ten independent simulations for survival times. A salient property of the survMNIST is that ground-truth cluster assignments cannot be identified based on the features or survival times alone; instead, a combination of the two has to be leveraged.

In addition to simulated data, we consider several real-world survival analysis problems. A popular benchmarking dataset utilised by previous works originates from the study to understand prognoses and preferences for outcomes and risks of treatments (**SUPPORT**) [252] on seriously ill adult patients at several tertiary care hospitals in the USA. The records include demographic, laboratory, and scoring data from subjects diagnosed with cancer, chronic obstructive pulmonary disease, congestive heart failure, cirrhosis, acute renal failure, multiple organ system failure, and sepsis.

Another publicly available benchmarking dataset we apply the methods to comes from the study investigating the association between serum free light chains (**FLChain**), proteins produced by plasma cells, and mortality. The data were acquired in Minnesota, USA, and include a few demographic and laboratory variables. Observe that the FLChain dataset is relatively low-dimensional (Table 4.1). Therefore, we expect little performance gain from more complex survival modelling approaches.

To further corroborate our findings, we perform experiments on a selection of in-house datasets. Among them, a cohort of patients treated surgically against high-grade glioma (**HGG**) [253] at the University Hospital Zürich, Switzerland, between 2008 and 2017. HGG is a type of malignant brain tumour, and this cohort includes two cancer types: glioblastoma and astrocytoma [254]. The dataset encompasses demographic, treatment, pre- and post-operative volumetric variables, information on the tumour location, histological findings, molecular markers, and performance scores. The cohort has few patients; thus, HGG serves as a benchmark representative of the "low data" regime.

A different challenge is posed by the dataset from patients undergoing chronic **hemodialysis** (HD) at DaVita Kidney Care (DaVita Inc., Denver, CO, USA) dialysis centres [255]–[259]. Hemodialysis is a therapy comprising blood filtering, typically applied in kidney patients. The current cohort includes patients who started HD under 18 years old and received it thrice per week between 2004 and 2016. The subjects were followed up on until the age of 30 years. The data represent an array of variables: demographics, disease etiology, and treatment-related information, such as the dialysis dose, fluid removal, and interdialytic weight gain. For a more detailed description of the data acquisition and variables, we refer the reader to the work by Gotta *et al.* [259]. The main challenge of the underlying prediction task is the high percentage of censored observations.

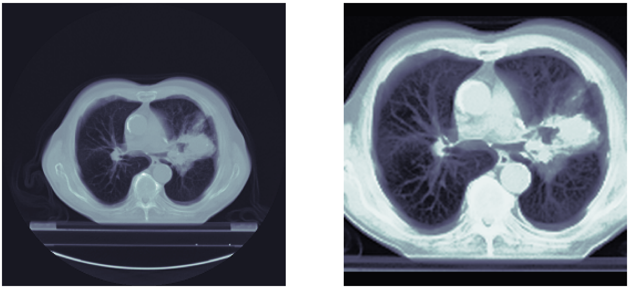
Lastly, as a high-dimensional unstructured data benchmark, we pool several cohorts of non-small cell lung cancer (**NSCLC**) patients. We focus on computed tomography (CT) scans and CT components of the positron emission tomography–computed tomography (PET/CT) scans acquired before treatment. We utilise several in-house and publicly available datasets:

- An in-house dataset [260] of PET/CT scans from 392 patients at the University Hospital Basel, Switzerland. The images are accompanied by tumour delineations, clinical data, and survival times.

- Lung1 dataset [261], [262] of CT scans from 422 patients at the Maastro Clinic, Maastricht, the Netherlands, alongside tumour segmentation and clinical and survival data. Lung1 and the datasets below are available from the Cancer Imaging Archive (TCIA) [263].

- Lung3 dataset [261], [264] of CT scans from 89 patients at the Maastro Clinic with segmentation, gene expression, and survival data.

- NSCLC Radiogenomics dataset [265]–[267] of CT and PET/CT scans from 211 patients at the Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System, USA.



(a) Original (b) Preprocessed

FIGURE 4.6: An example image from the Lung1 dataset [261], [262] produced by averaging 11 consecutive CT scan slices around the tumour (a) before and (b) after preprocessing.

We introduce several selection criteria and preprocessing steps to harmonise data from the different sources and reduce memory and computational complexity. In particular, we only retain data points with the *transversal* CT or PET/CT scan and tumour segmentation, excluding subjects with the largest transversal tumour area outside the lungs. Thus, the final pooled cohort includes 961 subjects (Table 4.1). Instead of directly utilising three-dimensional (3D) scans, we construct 2D images by averaging slices within 15 mm of the slice containing the tumour with a maximum transversal area. Consequently, images are normalised using histogram equalisation and downscaled to a resolution of 64×64 pixels. Additionally, we apply augmentations [268] to images during training to avoid spurious correlations. Figure 4.6 contains an example of a 2D image before (Figure 4.6a) and after (Figure 4.6b) preprocessing.

In summary, we conduct experiments on diverse datasets to explore different regimes and aspects of our model's behaviour. Specifically, we utilise synthetic benchmarks to evaluate clustering prediction. For real-world tabular datasets, we focus on the time-to-event prediction performance and low data and high censoring regimes. Finally, we demonstrate the method's applicability to the high-dimensional and unstructured NSCLC dataset and showcase the model's interpretability, providing in-depth qualitative results.

4.3.2 *Baselines and Ablations*

As briefly discussed in Section 4.1.1, the prior literature has explored many neural-network- and mixture-model-based approaches to nonlinear survival analysis. Throughout the experiments, we compare the VaDeSC model to several techniques within the scope of mixture modelling and clustering for survival analysis and unsupervised learning. In this subsection, we comment on the choice of baseline methods and ablation studies.

We use the semi-supervised clustering (SSC) technique introduced by Bair and Tibshirani [223] as a common-sense clustering baseline for survival data. As neural-network-based mixture modelling approaches, we consider survival cluster analysis by Chapfuwa *et al.* [228] and deep survival machines by Nagpal, Li, and Dubrawski [213]. SCA and DSM are closely related to the VaDeSC but assume distinct data-generating mechanisms and do not include a decoder neural network.

As a naïve unsupervised baseline, we apply k-means clustering in the raw feature space. We also perform ablations on the VaDeSC by (i) omitting the cluster assignments and mixture of Gaussians prior (VAE + Weibull) and (ii) removing the survival model (VaDE). The former variant corresponds to training a VAE [105] (Section 4.1.2) end-to-end with supervision from survival data, similar to the deep survival analysis model proposed by Ranganath *et al.* [209]. In this ablation, we apply k-means clustering *post hoc* to the latent space of the VAE. On the other hand, the latter variant is equivalent to the deep variational unsupervised clustering with VaDE [235]. Furthermore, recall that, in VaDeSC, cluster assignments can be made conditional on the survival time or without accounting for it (Equation 4.24). We report clustering results for both approaches.

Lastly, we utilise Cox PH and Weibull AFT models as simple survival time prediction baselines. When applying these conventional models to the NSCLC dataset, we extract radiomic features [269] from preprocessed CT images with the regions of interest given by tumour segmentation. Radiomics comprise automated feature extraction procedures to capture phenotypic characteristics in medical imaging data.

For a fair comparison, across all neural-network-based models, if possible, we utilise comparable encoder and decoder architectures and latent space dimensions. Appendix B.3 provides a detailed description of the architectures. Note that, for the NSCLC dataset, we use CNNs. When applicable, we set the number of mixture components to the ground truth and keep it the same for all finite mixture methods when the clustering structure is unknown. To reduce the computational cost, we use $L = 1$ MC samples (Equations 4.17–4.22) for our SGVB estimator.

4.3.3 *Evaluation Metrics*

To facilitate quantitative comparison, we resort to a few evaluation measures for clustering and time-to-event prediction performance. For clustering, we report the adjusted Rand index (ARI), normalised mutual information (NMI), and accuracy (ACC), computed using the Hungarian algorithm to find an optimal alignment of the assigned and ground-truth cluster labels.

For time-to-event prediction, we leverage a few measures to capture several performance aspects. Firstly, we assess the concordance index (C-index) [270], quantifying the model's ability to *rank* individuals w.r.t. risk. Let η_i for $1 \leq i \leq N$ be the predicted risk scores; then, the C-index is given by

$$\text{C-index} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}_{t_j<t_i}\mathbf{1}_{\eta_j>\eta_i}\delta_j}{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}_{t_j<t_i}\delta_j}. \tag{4.25}$$

Note that a C-index of 1 corresponds to the perfect ranking, whereas a random ranking of individuals is expected to have a C-index of 0.5. In practice, the predicted risk scores can be of an arbitrary scale. For instance, assuming the survival model described in Section 4.2.2, risk may be predicted as $\eta_i = 1/\lambda\left(z_i, c_i\right)$ for $1 \leq i \leq N$.

Another metric we consider is the relative absolute error (RAE) [271], corresponding to the average relative absolute deviation of the predicted from ground-truth survival time. As opposed to conventional regression analysis, we must account for censoring. For noncensored points, RAE is

$$\text{RAE}_{\text{nc}} = \frac{\sum_{i=1}^{N}\left|\left(\hat{t}_i - t_i\right)/t_i\right|\delta_i}{\sum_{i=1}^{N}\delta_i}, \tag{4.26}$$

where $\hat{t}_i$ is the predicted survival time for data point $1 \leq i \leq N$. Following the definition given by Chapfuwa *et al.* [228], for censored data points, the error is given by

$$\text{RAE}_{\text{c}} = \frac{\sum_{i=1}^{N}\left|\left(\hat{t}_i - t_i\right)/t_i\right|\left(1-\delta_i\right)\mathbf{1}_{\hat{t}_i\leq t_i}}{\sum_{i=1}^{N}\left(1-\delta_i\right)}. \tag{4.27}$$

Finally, we also assess if the estimated risk scores are reliable, i. e. well-calibrated. Calibration is an often overlooked aspect of predictive model performance [272]. In survival analysis, event rates predicted by a well-calibrated model should match the observed rates across time intervals [273]. Quantitatively, the calibration can be assessed by linearly regressing the predicted rates on the observed and inspecting the slope of the line [272]. An ideal model should have a calibration slope (CAL) of 1.0. Deviations from this value suggest systematic under- or overestimation of risk.

Most quantitative results are reported as averages alongside standard deviations computed across several simulations or using Monte Carlo cross-validation (CV). In addition to the metrics above, we explore qualitative results, including latent space embeddings, Kaplan–Meier (KM) curves [274], and prototype visualisations for individual clusters.

4.4 RESULTS

This section describes the experiment results. First, we focus on the datasets with known ground-truth clustering structures. Then, we turn to the time-to-event prediction on clinical tabular datasets followed by an in-depth exploration of our findings on real-world medical imaging data, showcasing our model's interpretability.

4.4.1 *Clustering*

As mentioned above, we first apply VaDeSC and baseline models to the synthetic benchmarks. Table 4.2 summarises clustering performance across the datasets. In addition to accuracies, ARIs, and NMIs, we report C-index values for time-to-event prediction as a sanity check, including the results for the Cox PH model as a simple baseline.

For both datasets, clustering in the raw feature space or on preselected covariates using k-means and SSC produces unsatisfactory results. Furthermore, the Cox-regression-based feature selection performed by the SSC yields little improvement over the conventional k-means. For survMNIST, these methods achieve a better absolute performance, nevertheless being less accurate than the rest of the techniques.

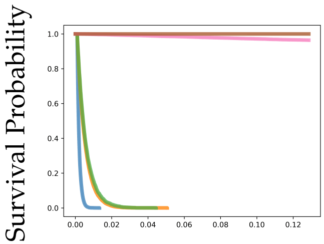
For the synthetic tabular data, whose generative process directly matches the assumptions of the VaDeSC, our model and VaDE outperform other neural-network-based techniques by a margin, including a VAE trained with the survival prediction loss but without the mixture prior. By contrast, on survMNSIST, the performance gap between VaDeSC and baselines is substantially smaller; however, on average, VaDeSC's cluster assignments,

Dataset	Method	ACC	NMI	ARI	C-index
Synthetic	*k*-means	0.44±0.04	0.06±0.04	0.05±0.03	—
	Cox PH	—	—	—	0.77±0.02
	SSC	0.45±0.03	0.08±0.04	0.06±0.02	—
	SCA	0.45±0.09	0.05±0.05	0.04±0.05	*0.82±0.02*
	DSM	0.37±0.02	0.01±0.00	0.01±0.00	0.76±0.02
	VAE + Weibull	0.46±0.06	0.09±0.04	0.09±0.04	0.71±0.02
	VaDE	0.74±0.21	0.53±0.12	0.55±0.20	—
	VaDeSC (w/o t)	*0.88±0.03*	*0.60±0.07*	*0.67±0.07*	**0.84±0.02**
	VaDeSC	**0.90±0.02**	**0.66±0.05**	**0.73±0.05**	
survMNIST	*k*-means	0.49±0.06	0.31±0.04	0.22±0.04	—
	Cox PH	—	—	—	0.74±0.04
	SSC	0.49±0.06	0.31±0.04	0.22±0.04	—
	SCA	0.56±0.09	0.46±0.06	0.33±0.10	*0.79±0.06*
	DSM	0.54±0.11	0.40±0.16	0.31±0.14	*0.79±0.05*
	VAE + Weibull	0.49±0.05	0.32±0.05	0.24±0.05	0.76±0.07
	VaDE	0.47±0.07	0.38±0.08	0.24±0.08	—
	VaDeSC (w/o t)	*0.57±0.09*	*0.51±0.09*	*0.37±0.10*	**0.80±0.05**
	VaDeSC	**0.58±0.10**	**0.55±0.11**	**0.39±0.11**	

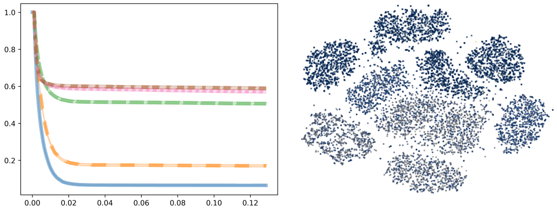
TABLE 4.2: Test-set clustering performance results on synthetic benchmarking datasets. For VaDeSC, we assess cluster assignment with and without survival time given as input (w/o t). For reference, we also report concordance index (C-index) values for time-to-event prediction. **Bold** indicates the best results, *italics* indicates the second best.

given or without survival time, are more coherent with the ground truth. Finally, we observe that SCA, DSM, and VaDeSC attain a comparable or higher C-index than that of the Cox PH model for both tasks.

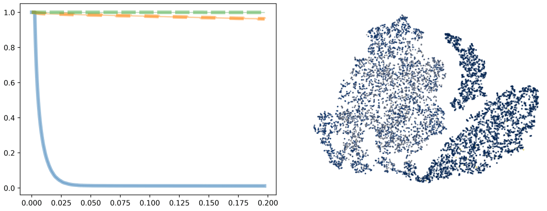
Let us now turn to qualitative findings from an exploration of the cluster assignments and latent embeddings inferred by the models. Figure 4.7 contains an overview of the results. Panels on the left show cluster-specific KM curves for the ground-truth structure (Figure 4.7a) and inferred assignments (Figures 4.7b–4.7e), whereas panels on the right depict 2D t-SNE [275] visualisations of the representations. As shown in Figure 4.7a, some survMNIST clusters have similar marginal survival distributions, and hence, both the covariates *and* survival times are necessary to identify the underlying clustering structure.



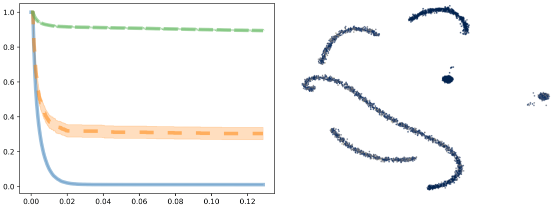
(a) Ground truth



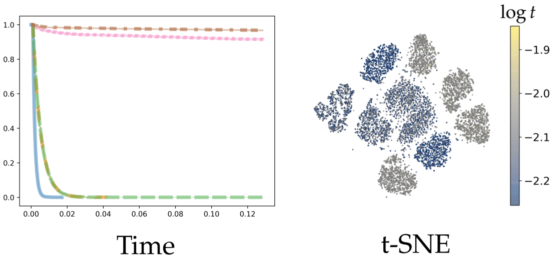
(b) VAE + Weibull



(c) SCA



(d) DSM



(e) VaDeSC

Figure 4.7: Qualitative results for clustering on the survMNIST data. The panels on the *left* show cluster-specific Kaplan–Meier (KM) curves. Herein, a separate curve is plotted for each cluster discovered by a method, and different colours correspond to individual clusters. The KM curves of the ground-truth clusters are shown in panel (a). Panels on the *right* contain t-SNE visualisations of representations learnt by the models. Data points are coloured by their respective observed survival times, with lower and higher values marked in **blue** and **yellow**.

The VAE model without the mixture-of-Gaussians prior (Figure 4.7b) learns clustered representations correlated with hand-written digits. However, cluster assignments produced by k-means performed *post hoc* on the latent space yield cluster-specific distributions different from the ground truth. Similarly, DSM (Figure 4.7d) produces compact clusters with disparate KM curves. Note that two clusters are empty despite initialising the model with five components. Among the baseline methods, SCA (Figure 4.7c) infers clusters with the closest correspondence to the underlying structure. Nevertheless, its embeddings are poorly correlated with digits, and only three clusters are discovered instead of five. Finally, VaDeSC discovers groups whose marginal survival distributions are most similar to the known structure (cf. Figure 4.7a), and its representations are visibly correlated with the observed time to event and digits.

4.4.2 *Time-to-event Prediction*

Beyond clustering structure discovery, another aspect we assess is survival time prediction. Below, we concentrate on clinical tabular datasets, while results for the more complex NSCLC data are described in Section 4.4.3.

Table 4.3 reports time-to-event prediction results w.r.t. C-index, relative absolute error for censored and noncensored data points, and calibration slope (Section 4.3.3). Generally, on these simpler datasets, neural-network-based approaches attain limited improvement over conventional Cox PH and Weibull AFT models, sometimes achieving lower RAEs and better calibration, e. g. on SUPPORT and Hemodialysis. On very small HGG and low-dimensional FLChain, expectedly, there are few differences among the methods. For Hemodialysis, we observe a relatively poor calibration across all models, which can be attributed to a very high percentage of censored observations (Table 4.1).

In summary, while our experiments do not suggest considerable predictive performance gains from more sophisticated modelling, overall, VaDeSC does not overfit on clinical tabular data, offering performance balanced across most metrics and comparable to closely related models, such as SCA and DSM. Given the characteristics of the datasets explored in this experiment, for instance, low dimensionality, data scarcity, and frequent censoring, the current findings are unsurprising and primarily serve as a proof of concept.

Dataset	**Method**	**C-index**	**RAE$_{\text{nc}}$**	**RAE$_{\text{c}}$**	**CAL**
SUPPORT	Cox PH	0.84±0.01	—	—	—
	Weibull AFT	0.84±0.01	0.62±0.01	*0.13±0.01*	*1.27±0.02*
	SCA	0.83±0.02	0.78±0.13	**0.06±0.04**	1.74±0.52
	DSM	**0.87±0.01**	*0.56±0.02*	*0.13±0.04*	1.43±0.07
	VAE + Weibull	0.84±0.01	*0.56±0.02*	0.20±0.02	1.28±0.04
	VaDeSC	*0.85±0.01*	**0.53±0.02**	0.23±0.05	**1.24±0.05**
FLChain	Cox PH	**0.80±0.01**	—	—	—
	Weibull AFT	**0.80±0.01**	*0.72±0.01*	**0.02±0.00**	*2.18±0.07*
	SCA	0.78±0.02	**0.69±0.08**	*0.05±0.05*	**1.33±0.24**
	DSM	*0.79±0.01*	0.76±0.05	**0.02±0.01**	2.35±0.66
	VAE + Weibull	**0.80±0.01**	0.76±0.01	**0.02±0.00**	2.55±0.07
	VaDeSC	**0.80±0.01**	0.76±0.01	**0.02±0.00**	2.52±0.08
HGG	Cox PH	*0.74±0.05*	—	—	—
	Weibull AFT	*0.74±0.05*	0.56±0.04	0.14±0.09	1.16±0.10
	SCA	0.63±0.08	0.97±0.05	**0.00±0.00**	2.59±1.70
	DSM	**0.75±0.04**	0.57±0.05	0.18±0.07	**1.09±0.08**
	VAE + Weibull	**0.75±0.05**	**0.52±0.06**	*0.12±0.07*	1.14±0.11
	VaDeSC	*0.74±0.05*	*0.53±0.06*	0.13±0.07	*1.12±0.09*
Hemodialysis	Cox PH	**0.83±0.04**	—	—	—
	Weibull AFT	**0.83±0.05**	0.81±0.03	**0.01±0.00**	*4.46±0.59*
	SCA	0.75±0.05	0.86±0.07	*0.02±0.02*	7.93±3.22
	DSM	*0.80±0.06*	0.85±0.08	*0.02±0.04*	8.23±4.28
	VAE + Weibull	0.77±0.06	*0.80±0.06*	*0.02±0.01*	4.49±0.75
	VaDeSC	*0.80±0.05*	**0.78±0.05**	**0.01±0.00**	**3.74±0.58**

TABLE 4.3: Test-set time-to-event prediction results across clinical tabular datasets.

4.4.3 *Non-small Cell Lung Cancer and Computed Tomography*

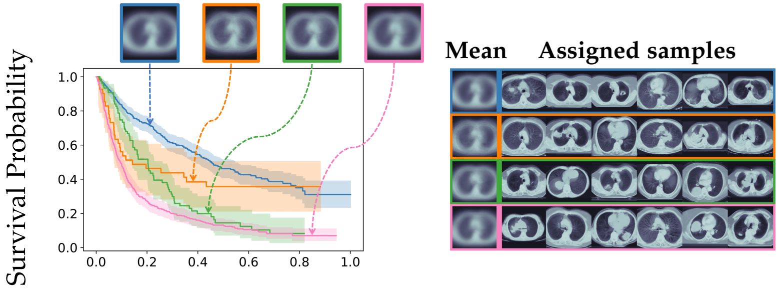
We now delve into a more complex application of the VaDeSC model to high-dimensional and unstructured data. The NSCLC dataset combines many practical challenges mentioned before: (i) dimensionality and input structure, (ii) small sample size, and (iii) heterogeneity owing to the pooling of multiple sources (Section 4.3.1). Beyond the predictive performance assessment, we leverage this dataset to demonstrate the interpretability and utility of our model in exploratory analysis. We also provide a more thorough empirical comparison with the deep survival machines, omitting the SCA, which previously attained results similar to DSM (Table 4.3).

Table 4.4 reports evaluation metrics for the survival time prediction. As conventional baselines, we consider Cox PH and Weibull AFT models fitted on features extracted from preprocessed images and tumour delineations [269]. Generally, neural-network-based DSM and VaDeSC perform comparably to the classical models trained on radiomics features. Note that, by contrast, neither DSM nor VaDeSC require laborious delineations since both models rely on representation learning instead of feature engineering.

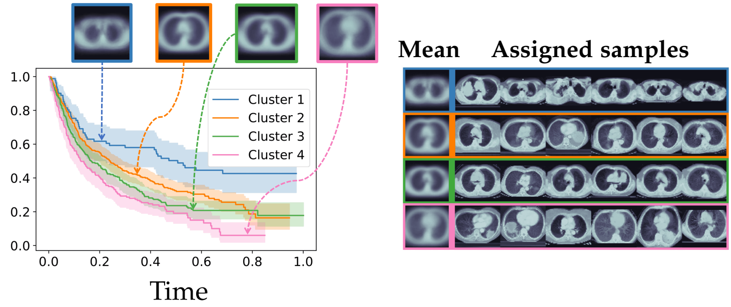
Figures 4.8–4.9 visualise clusters discovered by the two mixture models. In particular, Figure 4.8 shows cluster-specific KM curves alongside examples of CT images assigned to the groups and "centroids" computed by averaging. Note that we consider four components, as this configuration led to clustering structures consistent across the folds of the MC CV. Both techniques discover clusters with disparate survival distributions, as evidenced by KM curves. VaDeSC's clusters are correlated with tumour location reflected by CT images, as we observe systematic differences between centroids and assigned samples (Figure 4.8b). On the contrary, DSM's clusters have no clearly visible differences (Figure 4.8a).

Method	**C-index**	**RAE_{nc}**	**RAE_{c}**	**CAL**
Radiomics + Cox PH	**0.60±0.02**	—	—	—
Radiomics + Weibull AFT	**0.60±0.02**	**0.70±0.02**	0.45±0.03	1.26±0.04
DSM	*0.59±0.04*	0.72±0.03	**0.34±0.06**	*1.24±0.07*
VaDeSC	**0.60±0.02**	*0.71±0.03*	*0.35±0.05*	**1.21±0.05**

TABLE 4.4: Test-set time-to-event prediction results on NSCLC data. Cox PH and Weibull AFT models were trained on the radiomics features extracted from CT images with tumour segmentation.



(a) DSM



(b) VaDeSC

FIGURE 4.8: Clustering results on the NSCLC dataset for (a) DSM and (b) VaDeSC models. Panels on the *left* show Kaplan–Meier curves for the clusters discovered alongside corresponding "centroids" computed by averaging images assigned to each cluster. Panels on the *right* contain randomly chosen example images from every group.

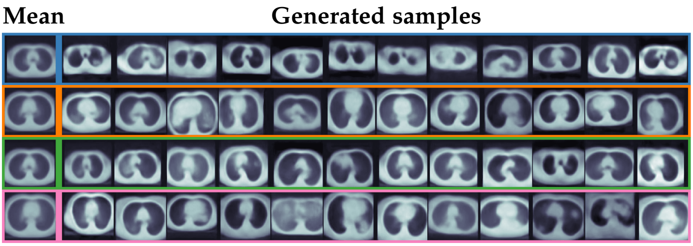


FIGURE 4.9: Images generated by VaDeSC for each of the clusters. To generate an image from a given cluster, we (i) sample a representation from the relevant component of the Gaussian mixture and (ii) map the representation to the image using the decoder. Mean samples on the *left* are generated by decoding the means of the Gaussian mixture.

Variable	**DSM**					**VaDeSC**				
	1	2	3	4	p-val.	1	2	3	4	p-val.
Tumour volume, cm^3	23	39	38	51	$\leq$ **1e-3**	43	36	40	63	$\leq$ **5e-2**
Age, yrs	67	68	68	69	0.11	62	69	67	70	$\leq$ **1e-3**
Female, %	29	30	26	21	0.3	36	19	38	23	$\leq$ **1e-3**
Smoker, %	84	100	80	89	0.9	67	94	87	100	0.12
Distant metsatasis, %	40	55	16	42	0.4	20	45	44	45	0.2
Tumour stage $\geq$ **3**, %	27	12	23	32	0.2	10	29	35	31	0.7

TABLE 4.5: Summary statistics, given by medians and percentages, for several demographic and clinical characteristics stratified by the cluster assignment made by DSM (*left*) and VaDeSC (*right*). In addition, we report p-values (p-val.) from the Kruskal–Wallis H test for the difference in medians across multiple independent groups.

Figure 4.9 demonstrates how VaDeSC can help interpret the relationship between features, learnt clusters, and survival times owing to its generative nature. Herein, we apply the decoder to the mean vectors of the mixture of Gaussians and conditionally generate samples from each component. We can then associate visualised decoded means and samples with cluster-specific survival distributions by inspecting KM curves (Figure 4.8b). Thus, clusters in VaDeSC effectively serve as prototype-based explanations of the nonlinear and high-dimensional association between images and survival. Similar to Figure 4.8, we observe a correlation between the tumour's location and cluster labels. Specifically, the lowest-risk cluster (1) is associated with the tumour in the upper section of the lungs. In contrast, the highest-risk cluster is characterised by the tumour in the lower section (4). This finding agrees with multiple previous analyses showing a higher five-year survival rate in NSCLC patients with upper-lobe tumours [276]. Notably, DSM fails to uncover such association, focusing on disparate risks (Figure 4.8a).

Additionally, we explore the association between clusters and extraneous clinically relevant characteristics [277]–[279], which were not explicitly included among input features. We consider the tumour stage and volume computed based on segmentation, patient's age, gender, smoking status, and presence of distant metastasis. Table 4.5 presents cluster-specific summary statistics for these variables. We also perform the Kruskal–Wallis H test [280] for differences in medians across clusters to assess if the discovered groups vary significantly w.r.t. the characteristics. The p-values indicate that DSM and VaDeSC infer clusters with significantly different tumour volumes. Across both models, the highest-risk group (4) has the largest median volume. Furthermore, unlike DSM, VaDeSC's clusters feature significant differences w.r.t. patients' age and gender. These results corroborate the qualitative findings above (Figures 4.8–4.9) and suggest that, overall, VaDeSC's clusters are driven by both covariates *and* survival times and vary w.r.t. a wider range of clinically relevant characteristics.

4.5 DISCUSSION

In this section, we summarise the contributions and findings of this chapter, discussing them in a broader context of the methods literature on survival analysis and interpretable machine learning. At the end, we comment on the methodological and practical limitations and potential improvements and directions for future research.

This chapter has treated survival analysis [206], a classical branch of biostatistics, with a focus on unstructured data, nonlinear relationships, and cluster analysis. Utilising previous works on variational autoencoders [105], deep variational clustering [235], and survival analysis [209], we have introduced a deep probabilistic model (Figures 4.4–4.5), VaDeSC, for clustering survival data. The model's parameters are optimised in a scalable and joint manner by maximising a lower bound on the joint likelihood of the observed data (Equation 4.16) using the SGVB estimator.

In addition to time-to-event prediction, the model allows uncovering clusters, or groups, of observations with variability in the relationship between the observed covariates and survival outcome. Moreover, VaDeSC also learns low-dimensional representations, or embeddings, that can be used in other downstream tasks. Since VaDeSC performs clustering and its latent space is regularised by the mixture of Gaussians prior, it allows producing prototype-based explanations (Section 2.4.4) for its time-to-event predictions. Similar to the generative Bayesian case model by Kim, Rudin, and Shah [138], a nonlinear relationship between the features and target variables can be elucidated using clusters and their quintessential exemplars, e. g. modes and centroids. Notably, every cluster is characterised by a simple relationship between the risk and latent space (Equation 4.10). Additional interpretation can be provided by visualising cluster-specific survival distributions using, for instance, the Kaplan–Meier estimator (Figure 4.8).

Beyond prototype- and case-based explanations, another aspect that makes VaDeSC more interpretable than other opaque neural-network-based

	SSC	**PR**	**SCA**	**DSM**	**VaDeSC**
Time-to-event prediction	✗	✓	✓	✓	✓
Representation learning	✗	✗	✓	✓	✓
Joint likelihood of x and t	✗	✓	✗	✗	✓
Scalability	✗	✗	✓	✓	✓
Adaptive # components	✗	✓	✓	✗	✗

TABLE 4.6: Comparison between VaDeSC and closely related survival analysis methods: semi-supervised clustering (SSC) [223], profile regression (PR) [230], survival cluster analysis (SCA) [228], and deep survival machines (DSM) [213]. We consider a few salient characteristics: whether (i) a model predicts the time to event, (ii) learns representations, (iii) maximises the joint likelihood of the observed covariates and survival times, (iv) is scalable to high-dimensional and unstructured data, and (v) does not require specifying a fixed number of clusters.

models [210]–[212] are explicit data-generating assumptions (Figure 4.5). Since VaDeSC is essentially a deep probabilistic graphical model [240], its assumptions w.r.t. (conditional) independences between variables are clearly formulated and can be scrutinised in the context of a specific application. Moreover, owing to this, the model can be adapted to alternative designs and settings if necessary.

As mentioned in Section 4.1.1, many related works have tackled the problem of clustering and mixture modelling of survival data. Table 4.6 compares VaDeSC to several closely related methods. Overall, our model provides a few crucial advantages. Semi-supervised clustering [223] is the most limited approach, as it relies on poorly scalable k-means and does not explicitly model time to event. While Bayesian profile regression assumes a generative process very similar to ours [230], it is not generalisable to unstructured datasets due to its reliance on MCMC methods. Lastly, SCA [228] and DSM [213] leverage neural networks, allowing for representation learning and survival time prediction. However, these models assume a different data-generating process, and their objectives result in purely outcome-driven clustering, as suggested by our experimental findings in Section 4.4. One limitation of VaDeSC is that, in contrast to the Dirichlet process mixture models [228], [230], it requires specifying a fixed number of clusters, which, in practice, is a nontrivial model selection problem [281].

A related *concurrent* work not included in Table 4.6 and experimental comparison (Section 4.4) is deep Cox mixtures [229]. Our approach differs from DCM in a few ways. DCM combines a VAE and a mixture of Cox PH regression models *empirically*, without deriving a valid evidence lower bound on the joint likelihood of x and t. Moreover, DCM relies on the computationally costly MC expectation–maximisation algorithm to infer cluster assignments instead of leveraging the mixture of Gaussians prior. Thus, despite some similarities in utilising VAEs and mixture modelling for survival analysis, the two approaches differ considerably in their implementation.

Last but not least, a distinctive feature of this chapter is the focus on *interpretable* survival analysis, specifically on prototype-based explanations that can be provided next to survival time predictions and cluster assignment inferred by VaDeSC. To the best of our knowledge, the related works discussed above do not explore this aspect of mixture models explicitly, concentrating purely on survival or cluster analysis.

In addition to methodological contributions, this chapter contains a comprehensive experimental evaluation of the proposed and related techniques w.r.t. time-to-event prediction and clustering (Sections 4.4.1–4.4.2). We observed that VaDeSC can identify clustering structures driven by both the covariates and survival outcomes. On the synthetic datasets with known cluster labels (Table 4.2), our method recovered the structure more accurately than baseline techniques. These findings were further corroborated by a closer inspection of cluster-specific survival distributions and low-dimensional embeddings (Figure 4.7).

We explored time-to-event predictions in more depth on several clinical tabular datasets (Table 4.3), demonstrating that VaDeSC attains competitive predictive performance in various scenarios, e. g. when data are scarce, low-dimensional, or few noncensored observations are available. Notably, our method exhibited a well-balanced performance w.r.t. concordance index, relative absolute error, and calibration.

We also showcased our model's utility on real-world medical imaging data (Section 4.4.3). In particular, we demonstrated that neural-network-based approaches can effectively predict survival time in NSCLC patients (Table 4.4) without extensive feature engineering and laborious tumour segmentation by medical experts. We observed that, by contrast to deep survival machines [213], VaDeSC learns clusters that, in addition to survival, are better correlated with phenotypic characteristics (Figure 4.8 and Table 4.5), such as the tumour's location and patient's age and gender. Lastly, we provided a detailed exploration of VaDeSC's interpretability by visualising prototype-based explanations alongside cluster-specific survival distributions (Figures 4.8–4.9). These explanations facilitate a better understanding of the nonlinear relationship between high-dimensional features and the model's time-to-event predictions.

A noteworthy empirical contribution of this chapter is the application of our technique to a real-world *image* dataset. Related works on the mixture modelling and clustering of survival data [213], [223], [228]–[230] have solely focused on tabular datasets with few covariates wherein neural networks bring limited performance gains at the cost of model opacity. Survival analysis on high-dimensional and unstructured data types, such as medical images [282], [283], is a problem practically pertinent to biomedical and healthcare domains and deserves more research attention.

4.5.1 *Limitations*

Our variational deep survival clustering framework features several conceptual and practical limitations discussed in the following. As mentioned above (Table 4.6), the choice of the number of mixture components is nontrivial and can be challenging to validate. While explicitly setting the number of clusters allows controlling the granularity of prototype-based explanations, it would be interesting to explore alternative model designs, for instance, Dirichlet process mixture models [228], [230], facilitating an *adaptive* number of components.

Prototypes make VaDeSC *locally* interpretable. However, the exact relationship between the covariates and predicted time to event remains opaque. This association could be additionally elucidated by simplifying the encoder neural network using a model design similar to that described in Chapter 3 in the context of time series analysis. Such modifications would further improve the model's interpretability, allowing us to evaluate the influence of individual covariates on the model's predictions.

Some adjustments to the experimental setup could be implemented to improve the quality of our empirical investigation. In particular, the experiments on the NSCLC dataset utilised extensive preprocessing (Section 4.3.1) to convert 3D CT scans into 2D images centred around the slice with the largest transversal tumour area. To eliminate remaining reliance on tumour segmentation and leverage all available information, we should explore the use of 3D convolutions as part of encoding and decoding neural networks.

Lastly, a natural limitation of VaDeSC is that, in specific application scenarios, its reliance on both covariate and survival information may be undesirable. If we seek to discover patient groups stratified purely by risk, for instance, for better-informed bedside management of severely ill patients, other techniques, e. g. DSM [213], might be more appropriate. By contrast, we expect VaDeSC to be helpful in the setting of *exploratory* analysis when applied to high-dimensional, unstructured, and heterogeneous datasets.

4.5.2 *Future Work*

As outlined in Section 4.5.1, several adjustments can be made to our model and experimental design. In addition, one general direction for future research is the exploration of more specialised settings and tasks, for instance, dynamic [284] and competing risks [211] survival analysis problems. Another aspect which deserves a deeper investigation is representation learning [104]. Although we briefly explored embeddings produced by the considered techniques in our experiments (Figure 4.7), additional insights could be gained from traversing the latent space of VaDeSC to understand if the model learns disentangled, i. e. more interpretable, representations [285]. From the generative modelling perspective, another noteworthy enhancement could come from the exploration of alternatives to VAEs, which are conveniently generalisable to various settings and assumptions, but have been long outperformed in density estimation and generated sample quality by more modern approaches. Therefore, a logical extension of this work would be the incorporation of other generative approaches, for example, variational diffusion models [286], hierarchical VAEs [287], or spatial dependency networks [288]. Such enhancements, however, may require nontrivial adjustments in our mixture modelling approach to discover meaningful clustering structures. To improve VaDeSC's practical utility, it could be helpful to generalise the model to multimodal datasets, e. g. building on previous advances in multimodal VAEs [243], [289]. Lastly, careful interpretation and validation of the structures discovered on all clinical datasets is beyond the scope of this thesis but forms a promising direction for future research.

4.6 SUMMARY

This chapter treated the problem of survival analysis, which frequently arises in the biomedical and healthcare domains. Survival analysis seeks to understand the association between observed covariates and the time to some (typically) adverse and clinically relevant event. The primary challenges of this task comprise censoring of the response variable, high dimensionality, and nonlinearity of the relationships.

Building on the previous efforts on deep generative and survival modelling, we introduced a probabilistic method to cluster high-dimensional, unstructured data accompanied by potentially censored survival times in a deep variational setting. Our model allows for (i) time-to-event prediction, (ii) discovery of covariate- and outcome-driven subpopulations, and (iii) representation learning. Owing to the regularisation of the model's latent space by the mixture of Gaussians prior, the predictions can be supplemented with prototype-based explanations by, for instance, visualising samples assigned to or conditionally generated from the same cluster as the data point of interest.

We conducted comprehensive experiments on synthetic and real-world clinical data and outlined conceptual arguments for the utility and novelty of the introduced technique. The results show that our method uncovers clusters correlated with both risk and phenotypic characteristics and has a competitive time-to-event prediction performance. Its ability to perform clustering and resulting prototype-based explanations are helpful for the exploratory analysis of high-dimensional and unstructured datasets.

5

CONCEPT-BASED MODELS IN THE WILD

Biomedical and healthcare datasets are often densely annotated, containing fine-grained labels related to the final target variable of interest. Such annotation can help design interpretable models reliant on *high-level* and human-understandable information. In Chapters 3 and 4, we have introduced models that provide some form of interpretation w.r.t. the feature space. However, in some tasks and settings, this space may be altogether uninterpretable or too high-dimensional, e. g. think of high-throughput multiview or multimodal medical image examination. In such cases, we may resort to high-level, typically categorical, variables succinctly describing the object's features and bearing a close relation to the response. These variables are usually understandable to a domain expert and are referred to as *attributes* or *concepts*. Recent interpretable and explainable ML literature has seen a renewed interest in concept-based models and explanation techniques (Sections 2.3.4 and 2.4.3).

A classic example of the model class that utilises attributes is concept bottleneck models (Figure 2.3c) [22], [111], [112], which comprise step-by-step prediction of (i) the attributes from the features and (ii) the target variable from the predicted attribute values. This chapter will focus on CBMs and their scalability to more complex datasets and application scenarios. In particular, motivated by real-world medical imaging applications, we enhance CBMs and generalise them to classification problems involving *multiview* and *multimodal* data [32], [33]. Furthermore, we tackle the challenge of *systematically* missing concept variables unobservable due to the lack of domain knowledge, impossibility or unethicality of measurement, for instance, when the assessment of concepts requires a costly or invasive procedure.

Another contribution of this chapter is the application of the mentioned enhanced CBMs to predict the diagnosis and disease severity and guide the management of pediatric patients admitted with suspected appendicitis to an emergency department based on multiview ultrasound imaging data. Thus, we demonstrate that our approach is effective "in the wild", helping develop accurate and interpretable classifiers. From the clinical perspective, the application we investigate is highly relevant, as appendicitis is one of the most frequent causes of abdominal pain resulting in hospital admissions in children [290] and effective triage of patients with suspicion of this disease remains a sought-after "holy grail" of pediatric surgical research [291].

In the following sections, we describe the broader context and prior works related to this chapter, introduce enhanced CBMs to handle multiview data under incomplete concepts, explain the experimental setup and describe and discuss our findings. This chapter is based on the contents and text of the peer-reviewed article "Interpretable and Intervenable Ultrasonography-based Machine Learning Models for Pediatric Appendicitis" [292].

5.1 BACKGROUND

To outline a broader context of the current chapter, this section describes related works on concept-based modelling, multiview and multimodal learning, and applications of ML to appendicitis-related data.

To begin with, we provide a more formal description of the problem setting and introduce essential design considerations. We consider a dataset comprising triples $\left(\left\{\boldsymbol{x}_i^v\right\}_{v=1}^{V_i}, \boldsymbol{c}_i, y_i\right)$ for $1 \leq i \leq N$, where $\left\{\boldsymbol{x}_i^v\right\}_{v=1}^{V_i}$ is a sequence of view-specific features, $\boldsymbol{c}_i \in \mathbb{R}^K$ is a vector of K concepts, and y_i is the label. Note that the number of views $V_i \geq 1$ may vary across data points $1 \leq i \leq N$. We assume that all views can be preprocessed and rescaled into the same dimensionality. Nevertheless, our proposed methods can be readily adjusted to handle heterogeneous multimodal data.

Motivated by the properties of medical image datasets, this chapter assumes a few characteristics described informally below. (i) Firstly, not every concept variable may be identifiable from *each* view (*partial observability*), i. e. some concepts may be visible in few images. (ii) Secondly, we assume that views share a considerable amount of information, being visually and semantically similar (*view homogeneity*). (iii) Lastly, views within the same data point may be loosely ordered, e. g. spatially, temporally, or based on their importance for predicting the label (*view ordering*).

5.1.1 *Concept-based Models*

In Sections 2.3.4 and 2.4.3, we have briefly touched on concept-based models and explanation techniques. Since this research direction is the subject of the current chapter, we recapitulate and expand on the previous discussion.

Computer vision research has long studied the use of high-level attributes in predictive models [111], [112]. More recent works explore explicitly incorporating concepts in neural networks [22], [114], producing high-level *post hoc* explanations by quantifying the network's sensitivity to the attributes [129], probing [293], [294] and de-correlating and aligning the network's latent space with concept variables [113].

Below, we describe concept bottleneck models [22] more formally. In brief, a CBM $f_{\boldsymbol{\theta}}$ parameterised by $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\psi}\}$ is given by

$$f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = g_{\boldsymbol{\psi}}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right), \tag{5.1}$$

where $h_{\boldsymbol{\phi}}$ maps inputs to predicted concepts $\hat{\boldsymbol{c}} = h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)$ and $g_{\boldsymbol{\psi}}$ predicts the target based on $\hat{\boldsymbol{c}}$, i. e. $\hat{y} = g_{\boldsymbol{\psi}}\left(\hat{\boldsymbol{c}}\right)$. CBMs are usually trained on labelled data $\left\{\left(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i\right)\right\}_i$ by minimising concept and target prediction losses jointly, sequentially, or independently. Observe that, in Equation 5.1, $h_{\boldsymbol{\phi}}$ forms a *concept bottleneck layer*, and thus, the final output depends on the covariates $\boldsymbol{x}$ solely through the predicted concept values $\hat{\boldsymbol{c}}$. Throughout this chapter, we will refer to $h_{\boldsymbol{\phi}}$ and $g_{\boldsymbol{\psi}}$ as the concept and target models, respectively.

CBMs are deemed interpretable since concept predictions $\hat{\boldsymbol{c}}$ can be inspected alongside the final output $\hat{y}$ and utilised as concept-based explanations. Furthermore, in contrast to conventional multitask learning [295], we can intervene on and interact with the model at test time by editing concept predictions and affecting downstream output. For instance, if we choose to replace $\hat{\boldsymbol{c}}$ with another $\boldsymbol{c}'$, the final prediction must be updated to $\hat{y}' = g_{\boldsymbol{\psi}}\left(\boldsymbol{c}'\right)$. This process of editing the model's intermediate output is referred to as *intervention* and is a distinctive advantage of CBMs over other interpretable model classes, e. g. SENNs (Chapter 3).

Interventions facilitate human–model interaction and allow for the injection of the expert's knowledge. For example, the simplest intervention strategy is to "correct" the model, replacing predicted concept values with the ground truth. Let $\mathcal{S} \subseteq \{1, \ldots, K\}$ be a subset of concept variables to be intervened on, then, under this simple strategy, the updated prediction is

$$\hat{y}' = g_{\boldsymbol{\psi}}\left(\hat{\boldsymbol{c}}_{\{1,\ldots,K\}\setminus\mathcal{S}}, \boldsymbol{c}_{\mathcal{S}}\right), \tag{5.2}$$

where $\boldsymbol{c}$ is the ground-truth concept vector. Note the notation abuse in the order of the arguments in $g_{\boldsymbol{\psi}}$.

5.1.2 *Multiview and Multimodal Learning*

As evident from the previous subsection, vanilla concept bottleneck models (Equation 5.1) generally assume a single unimodal feature set. By contrast, medical imaging gives natural rise to *multiview* and *multimodal* data, e. g. in medical ultrasound (US) [296], [297]. For instance, the risk of breast cancer is routinely assessed based on multiview and multimodal US images of lesions, including transversal and longitudinal views of B(rightness)-mode, colour Doppler and elastography images.

Generally, multiview learning [32] concerns itself with the data comprising multiple views, essentially feature subsets, representing the source object. Similarly, multimodal learning [33] studies models combining, or fusing, multiple heterogeneous modalities, e. g. images and text. Beyond the supervised learning setting, both directions have seen advances in self-supervised [298], [299] and generative modelling approaches [300].

5.1.3 *Machine Learning for Appendicitis*

As this chapter's primary application of interest is predictive modelling for pediatric appendicitis, we provide basic background on the disease and an overview of related works leveraging machine learning methods.

The diagnosis of patients with suspected appendicitis can be challenging and relies on a combination of clinical, laboratory, and imaging parameters [301]. Despite extensive research, no specific and practically useful biomarkers for the early detection of appendicitis have been identified [302], [303]. Epidemiologically and clinically, there are two forms of appendicitis: uncomplicated (subacute/exudative, phlegmonous) and complicated (gangrenous, perforated) [303]–[305]. Management forms include surgery as the standard method [301], [306] and conservative therapy [304], [306]–[309].

Conventional imaging modalities for suspected appendicitis are ultrasonography, magnetic resonance imaging, and computed tomography. US has become the primary choice due to widespread availability, lack of radiation, and improvements in resolution over the past years [310]. Repeated US examinations, including B-mode and Doppler, during the observation phase improve diagnostic accuracy and help identify disease progression [306], [311], [312].

There is an abundance of works tackling the prediction of the diagnosis and management in pediatric and adult patients [313]–[322]. Most models either utilise simple clinical and laboratory data [313], [316], [317], [322], rely on hand-crafted US annotations [315], [318], [320], [321], or require more expensive and invasive imaging modalities, such as CT [319]. Despite having lower sensitivity and specificity than CT, US has been advocated as the preferred diagnostic modality due to the absence of ionising radiation and cost-effectiveness [323]. However, fully automated analysis of abdominal US images in this context remains an under-explored approach.

5.2 MULTIVIEW CONCEPT BOTTLENECK MODELS

Following the notation introduced at the beginning of Section 5.1, we generalise concept bottleneck models [22] (Section 5.1.1) to the multiview classification scenario. We will refer to this enhanced model as the *multiview concept bottleneck model* (MVCBM). The remainder of this section explains its modules and training procedure in detail.

5.2.1 *Model Architecture*

In brief, MVCBM consists of four modules: (i) per-view feature extraction, (ii) feature fusion, (iii) concept prediction, and (iv) label prediction. Figure 5.1 provides a schematic overview of the MVCBM's architecture, while the forward pass is specified by Equations 5.3a–5.3d below.

For data point $1 \leq i \leq N$, a forward pass of the MVCBM is given by the following equations:

I Feature extraction:

$$\boldsymbol{h}_i^{c,v} = h_{\boldsymbol{\omega}^c}\left(\boldsymbol{x}_i^v\right), 1 \leq v \leq V_i, \tag{5.3a}$$

II Feature fusion:

$$\bar{\boldsymbol{h}}_i^c = r_{\boldsymbol{\xi}^c}\left(\left\{\boldsymbol{h}_i^{c,v}\right\}_{v=1}^{V_i}\right), \tag{5.3b}$$

III Concept prediction:

$$\hat{\boldsymbol{c}}_i = s_{\boldsymbol{\chi}^c}\left(\bar{\boldsymbol{h}}_i^c\right), \tag{5.3c}$$

IV Label prediction:

$$\hat{y}_i = g_{\boldsymbol{\psi}}\left(\hat{\boldsymbol{c}}_i\right), \tag{5.3d}$$

where Latin letters correspond to functions and variables and Greek letters denote learnable parameters. Observe that parameters $\boldsymbol{\phi}^c = \{\boldsymbol{\omega}^c, \boldsymbol{\xi}^c, \boldsymbol{\chi}^c\}$ define the concept model $h_{\boldsymbol{\phi}^c}$ (Equation 5.1) mapping a multiview feature sequence to the predicted concept values, whereas $g_{\boldsymbol{\psi}}$ is the target model, linking the concepts and labels. Thus, similar to the vanilla concept bottleneck, MVCBM's forward pass can be rewritten as $\hat{y}_i = g_{\boldsymbol{\psi}}\left(h_{\boldsymbol{\phi}^c}\left(\left\{\boldsymbol{x}_i^v\right\}_{v=1}^{V_i}\right)\right)$. In the following, we describe every step from Equation 5.3.

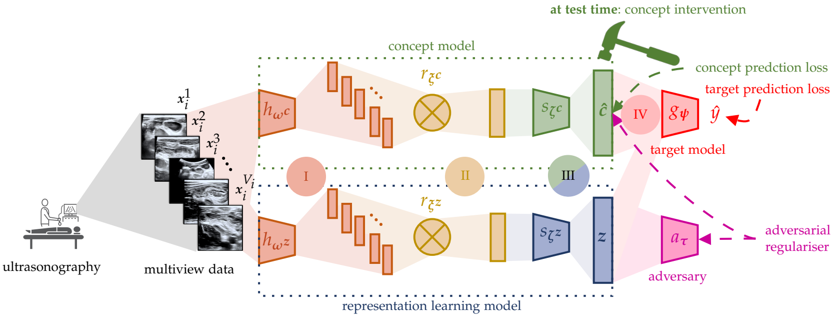


FIGURE 5.1: Schematic summary of the multiview concept bottleneck model (MVCBM) and its semi-supervised extension (SSMVCBM). (I) Multiview ultrasound images $\left(\{\boldsymbol{x}_i^v\}_{v=1}^{V_i}\right)$ are mapped to features using a shared encoder neural network $(h_{\omega^c}, h_{\omega^z})$; (II) features are aggregated across the views; (III) high-level human-understandable concepts and representations are predicted based on the aggregated features; (IV) using concepts and representations, the target prediction is made. The MVCBM only includes view encoding, fusion, and concept prediction, whereas the SSMVCBM also performs representation learning. During training, in addition to the *target prediction loss*, the MVCBM is supervised by the *concept prediction loss*. The SSMVCBM is further penalised by an *adversarial regulariser* encouraging statistical independence between predicted concepts and representations.

❶ The first step in the MVCBM's forward pass is **feature extraction**. Given an ordered view sequence $\{\boldsymbol{x}_i^v\}_{v=1}^{V_i}$, we encode each view into a lower-dimensional representation (Equation 5.3a). To this end, we employ a *shared* encoder network denoted by h_{ω^c}. Weight sharing is justified by view homogeneity (Section 5.1) and could be helpful in smaller datasets with missing features. By contrast, in multimodal datasets, dissimilarities across images acquired from the same subject are significant and consistent. In this scenario, it may be prudent to train a dedicated encoder for each modality to learn modality-specific features. In practice, we utilise a pretrained model to initialise the weights of h_{ω^c} [324].

II Having obtained a sequence of view-specific features $\left\{\boldsymbol{h}_i^{c,v}\right\}_{v=1}^{V_i}$, we perform **fusion**, or aggregation, as shown in Equation 5.3b. Following the *hybrid fusion* approach [33], we aggregate intermediate view-specific features $\boldsymbol{h}_i^{c,v}$ from the previous step within a single neural network instead of concatenating views at the input level (*early fusion*) or training an ensemble of view-specific models (*late fusion*).

Thus, the class of the fusion function r_{ζ^c} is one of the design choices behind our model. Although many function classes are viable, in the context of multiview medical imaging data, the fusion must handle varying numbers of views per data point. As a naïve approach, we consider the arithmetic mean across the views $\bar{\boldsymbol{h}}_i^c = \frac{1}{V_i}\sum_{v=1}^{V_i}\boldsymbol{h}_i^{c,v}$ [325], where $\bar{\boldsymbol{h}}_i^c$ denotes the fused feature vector. Considering partial observability of the concepts and ordering of the views, we also investigate aggregation via a *learnable* function. Similar to Ma *et al.* [326], who utilise this trick in multiview 3D shape recognition, we combine view-specific representations via an LSTM network [177].

III, **IV** Analogous to the vanilla CBM, the last two steps (Equations 5.3c–5.3d) are **concept and label prediction**. First, we predict concepts $\hat{\boldsymbol{c}}_i$ based on the fused representation $\bar{\boldsymbol{h}}_i^c$ using the network s_{χ^c}. The vector $\hat{\boldsymbol{c}}_i$ is then used as an input to the target model g_{ψ}, predicting the label $\hat{y}$.

5.2.2 *Loss Function and Optimisation*

We now define loss functions and procedures to be utilised for the optimisation of MVCBM's parameters. Recall that vanilla CBMs can be optimised using independent, sequential, and joint procedures [22]. This chapter will focus on the sequential and joint approaches that offer a more balanced trade-off between predictive performance and intervention effectiveness, as shown experimentally by Koh *et al.* [22].

For **sequential training**, we first optimise the concept model parameters:

$$\hat{\boldsymbol{\phi}}^c = \arg\min_{\boldsymbol{\phi}^c} \sum_{i=1}^{N}\sum_{k=1}^{K} w_i^y w_i^{c_k} \ell^{c_k}\left(\hat{c}_{i,k}, c_{i,k}\right), \tag{5.4}$$

where ℓ^{c_k} is the loss function for the k-th concept, e. g. the CE for categorical and MSE for numerical concepts, and $c_{i,k}$ refers to the value of the k-th concept for the i-th data point.

To address imbalances in concept label distributions and scarcity of specific concept-target combinations, we introduce weights $w_i^{c_k}$ and w_i^y for the k-th concept and target variable values in the i-th point. In practice, we set these weights to the normalised inverse counts of samples in the corresponding variable classes, i. e. $w_i^y \propto 1/\sum_{j=1}^{N} \mathbf{1}_{\{y_j=y_i\}}$ and $w_i^{c_k} \propto 1/\sum_{j=1}^{N} \mathbf{1}_{\{c_{j,k}=c_{i,k}\}}$. Notably, other sample weighting schemes may be viable.

In the next step of sequential training, parameters $\hat{\boldsymbol{\phi}}^c$ from Equation 5.4 are frozen, and the parameters of the target model $g_{\boldsymbol{\psi}}$ are optimised:

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} \sum_{i=1}^{N} w_i^y \ell^y \left(g_{\boldsymbol{\psi}}\left(\hat{\boldsymbol{c}}_i\right), y_i\right), \tag{5.5}$$

where ℓ^y is the loss function for the target prediction task and $\hat{\boldsymbol{c}}_i$ are predictions made by the frozen concept model $h_{\hat{\boldsymbol{\phi}}^c}$.

By contrast, **joint training** combines the loss functions from Equations 5.4 and 5.5 into a single objective:

$$\hat{\boldsymbol{\phi}}^c,\, \hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\phi}^c, \boldsymbol{\psi}} \left\{ \sum_{i=1}^{N} w_i^y \ell^y(\hat{y}_i, y_i) + \alpha \sum_{i=1}^{N} \sum_{k=1}^{K} w_i^y w_i^{c_k} \ell^{c_k}(\hat{c}_{i,k}, c_{i,k}) \right\}, \tag{5.6}$$

where $\alpha > 0$ controls the trade-off between the target and concept losses. As the name of this procedure suggests, parameters $\boldsymbol{\phi}^c$ and $\boldsymbol{\psi}$ are optimised simultaneously.

5.2.3 *Extension to Unobserved Concepts*

Vanilla CBMs and MVCBMs implicitly assume that concept variables are *complete* in that they fully capture the predictive relationship between the covariates and the target. This assumption may be false for practical reasons, such as the high cost of annotation, lack of knowledge, or ethical concerns regarding the measurement of certain variables. More formally, concept bottlenecks require that concepts are a *sufficient statistic* for the target variable [135]: $\boldsymbol{x} \perp\!\!\!\perp y \mid \boldsymbol{c}$. In violation of this assumption, conditional dependencies may occur when some ground-truth concept variables are systematically missing in the acquired dataset, i. e. unobserved for *all* data points. Figure 5.2 depicts non-exhaustive examples of data-generating mechanisms that may lead to the scenario described above. In such cases, the predictive performance of the CBM is limited since the model solely relies on the

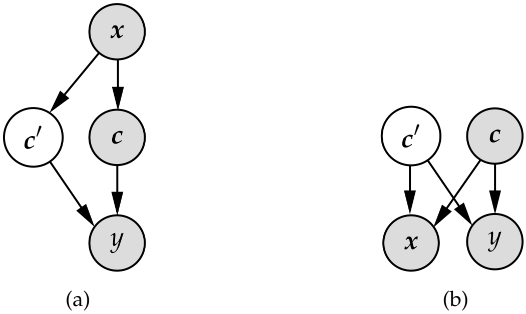


Figure 5.2: Generative processes with incomplete concept sets summarised as directed graphical models. Shaded and unshaded nodes correspond to observed and unobserved variables, respectively. For both (a) and (b), in general, $\boldsymbol{x} \not\perp\!\!\!\perp y \,|\, \boldsymbol{c}$ since there exists an active path between $\boldsymbol{x}$ and y through unobserved concepts $\boldsymbol{c}'$.

predefined set of concepts, which is insufficient. To this end, we propose a *semi-supervised* variant of the MVCBM (SSMVCBM) that additionally learns representations complementary to the concepts and relevant to the downstream prediction task.

Next to the feature extraction and concept prediction, SSMVCBM includes an unsupervised module mapping views $\left\{\boldsymbol{x}_i^v\right\}_{v=1}^{V_i}$ to the representation $\boldsymbol{z}_i \in \mathbb{R}^J$ (Figure 5.1). Analogous to Equations 5.3a–5.3c, the architecture comprises the following steps for data point $1 \leq i \leq N$: (i) $\boldsymbol{h}_i^{z,v} = h_{\boldsymbol{\omega}^z}\left(\boldsymbol{x}_i^v\right)$ for view $1 \leq v \leq V_i$, (ii) $\bar{\boldsymbol{h}}_i^z = r_{\boldsymbol{\xi}^z}\left(\left\{\boldsymbol{h}_i^{z,v}\right\}_{v=1}^{V_i}\right)$, and (iii) $\boldsymbol{z}_i = s_{\boldsymbol{\chi}^z}\left(\bar{\boldsymbol{h}}_i^z\right)$. Subsequently, for the final prediction, $\hat{\boldsymbol{c}}_i$ and $\boldsymbol{z}_i$ are concatenated and fed into the target model: for data point $1 \leq i \leq N$, $\hat{y}_i = g_{\boldsymbol{\psi}}\left(\left[\hat{\boldsymbol{c}}_i, \boldsymbol{z}_i\right]\right)$. As a shorthand notation, let $h_{\boldsymbol{\phi}^z}$ denote the entire representation learning model with parameters $\boldsymbol{\phi}^z = \left\{\boldsymbol{\omega}^z, \boldsymbol{\xi}^z, \boldsymbol{\chi}^z\right\}$.

Observe that the SSMVCBM model is semi-supervised in that the label is predicted based on both $\hat{\boldsymbol{c}}_i$ and $\boldsymbol{z}_i$, where $\hat{\boldsymbol{c}}_i$ are supervised by the concept prediction loss, while $\boldsymbol{z}_i$ are complementary representations learnt without explicit labels. These representations are meant to capture the residual relationship between $\boldsymbol{x}$ and y not represented among the observed concepts.

To avoid learning representations redundant to the concepts, we deem it desirable that $\hat{\boldsymbol{c}} \perp\!\!\!\perp \boldsymbol{z} \,|\, y$, i.e. the predicted concepts and representations should be statistically independent conditional on the label. To this end, we use another neural network $a_{\boldsymbol{\tau}} : \mathbb{R}^J \rightarrow \mathbb{R}^K$, parameterised by weights $\boldsymbol{\tau}$, to quantify the degree of statistical dependence as $\max_{\boldsymbol{\tau}} \operatorname{corr}\left(a_{\boldsymbol{\tau}}\left(\boldsymbol{z}\right), \hat{\boldsymbol{c}}\right)$

[327]. Thus, network a_τ is used to adversarially regularise representation $\boldsymbol{z}$. Empirically, we observed that this regularisation scheme helps de-correlate $\boldsymbol{z}$ from concept predictions and improves the effectiveness of interventions. Lastly, note that, in the context of the data-generating mechanisms shown in Figure 5.2, $\boldsymbol{z}$ does not need to identify unobserved concepts $\boldsymbol{c}'$ but rather represents the residual relationship between $\boldsymbol{x}$ and y.

In practice, we train SSMVCBMs using a specially tailored procedure outlined in Algorithm 2. Similar to the sequential optimisation for (MV)CBMs (Equations 5.4 and 5.5), it comprises multiple steps. First, parameters $\boldsymbol{\phi}^c = \{\boldsymbol{\omega}^c, \boldsymbol{\xi}^c, \boldsymbol{\chi}^c\}$ involved in concept prediction are optimised using the loss function from Equation 5.4 (lines 1–6). Then, we freeze $\hat{\boldsymbol{\phi}}^c$ and optimise representation learning model parameters $\boldsymbol{\phi}^z = \{\boldsymbol{\omega}^z, \boldsymbol{\xi}^z, \boldsymbol{\chi}^z\}$ (lines 7–19):

$$\hat{\boldsymbol{\phi}}^z, \tilde{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\phi}^z, \boldsymbol{\psi}} \max_{\tau} \sum_{i=1}^{N} w_i^y \ell^y\left(\hat{y}_i, y_i\right) - \lambda \sum_{i=1}^{N} \sum_{k=1}^{K} w_i^{c_k} \ell^{c_k}\left(\left[a_\tau\left(\boldsymbol{z}_i\right)\right]_k, \hat{c}_{i,k}\right), \quad (5.7)$$

where $\lambda > 0$ controls the weight of the adversarial regulariser. The loss function above can be extended with additional regularisation terms, e. g. to de-correlate individual dimensions of $\boldsymbol{z}$ [328], improving the interpretability of representations. The minimax objective from Equation 5.7 is optimised using an adversarial training technique similar to the one utilised for generative adversarial networks [329]. Last but not least, the parameters of the target prediction model are additionally re-optimised (cf. Equation 5.5), treating $\hat{\boldsymbol{\phi}}^c$ and $\hat{\boldsymbol{\phi}}^z$ as fixed (lines 20–25):

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} \sum_{i=1}^{N} w_i^y \ell^y\left(g_{\boldsymbol{\psi}}\left(\left[\hat{\boldsymbol{c}}_i, \boldsymbol{z}_i\right]\right), y_i\right). \quad (5.8)$$

From the conceptual perspective, the last step is not necessary, as the target model parameters are optimised as part of Equation 5.7. However, we observed that, with this step, the model's predictive performance becomes less sensitive to tuning the λ-parameter.

In summary, the semi-supervised variant of MVCBM tackles prediction problems under systemically missing concept variables (Figure 5.2) by including a representation learning branch (Figure 5.1) while retaining the interpretability and capability to handle multiview data from the ground model. Nevertheless, this enhancement comes at the cost of additional learnable and tuning parameters and a less stable and more computationally costly adversarial training procedure (Algorithm 2).

Algorithm 2: Training Procedure for SSMVCBM

```
Input: Training set $\mathscr{D}_{\text{train}} = \left\{ \left( \{ \boldsymbol{x}_i^v \}_{v=1}^{V_i}, \boldsymbol{c}_i, y_i \right) \right\}_{i=1}^{N}$; numbers of epochs
       $E_{\boldsymbol{c}}, E_{\boldsymbol{z}}, E_a, E_y \geq 1$ for the concept, representation learning, adversary, and target
       model optimisation; learning rates $\eta_{\boldsymbol{c}}, \eta_{\boldsymbol{z}}, \eta_a, \eta_y > 0$; number of adversarial
       training steps $C \geq 1$; regularisation parameter $\lambda \geq 0$
Output: Optimised SSMVCBM parameters $\{ \hat{\boldsymbol{\phi}}^c, \hat{\boldsymbol{\phi}}^z, \hat{\boldsymbol{\psi}} \}$

Initialise $\hat{\boldsymbol{\phi}}^c = \left\{ \hat{\omega}^c, \hat{\xi}^c, \hat{\chi}^c \right\}$
for $e = 1$ to $E_{\boldsymbol{c}}$ do
    for minibatch $\mathscr{B} \subseteq \{1, \ldots, N\}$ do
        Update $\hat{\boldsymbol{\phi}}^c \leftarrow \hat{\boldsymbol{\phi}}^c - \eta_c \nabla_{\hat{\boldsymbol{\phi}}^c} \sum_{i \in \mathscr{B}} \sum_{k=1}^{K} w_i^y w_i^{c_k} \ell^{c_k} \left( \hat{c}_{i,k}, c_{i,k} \right)$
    end
end

Initialise $\hat{\boldsymbol{\phi}}^z = \left\{ \hat{\omega}^z, \hat{\xi}^z, \hat{\chi}^z \right\}$, $\hat{\boldsymbol{\psi}}$, and $\hat{\boldsymbol{\tau}}$
for $j = 1$ to $C$ do
    for $e = 1$ to $E_{\boldsymbol{z}}$ do
        for minibatch $\mathscr{B} \subseteq \{1, \ldots, N\}$ do
            Update
                $\{ \hat{\boldsymbol{\phi}}^z, \hat{\boldsymbol{\psi}} \} \leftarrow \{ \hat{\boldsymbol{\phi}}^z, \hat{\boldsymbol{\psi}} \} - \eta_z \nabla_{\{ \hat{\boldsymbol{\phi}}^z, \hat{\boldsymbol{\psi}} \}} \Big[ \sum_{i \in \mathscr{B}} w_i^y \ell^y \left( \hat{y}_i, y_i \right)$
                $\qquad - \lambda \sum_{i \in \mathscr{B}} \sum_{k=1}^{K} \ell^{c_k} \left( \left[ a_{\hat{\boldsymbol{\tau}}} \left( \boldsymbol{z}_i \right) \right]_k, \hat{c}_{i,k} \right) \Big]$
        end
    end
    for $e = 1$ to $E_a$ do
        for minibatch $\mathscr{B} \subseteq \{1, \ldots, N\}$ do
            Update $\hat{\boldsymbol{\tau}} \leftarrow \hat{\boldsymbol{\tau}} - \eta_a \nabla_{\hat{\boldsymbol{\tau}}} \sum_{i \in \mathscr{B}} \sum_{k=1}^{K} w_i^{c_k} \ell^{c_k} \left( \left[ a_{\hat{\boldsymbol{\tau}}} \left( \boldsymbol{z}_i \right) \right]_k, \hat{c}_{i,k} \right)$
        end
    end
end

Reinitialise $\hat{\boldsymbol{\psi}}$
for $e = 1$ to $E_y$ do
    for minbatch $\mathscr{B} \subseteq \{1, \ldots, N\}$ do
        Update $\hat{\boldsymbol{\psi}} \leftarrow \hat{\boldsymbol{\psi}} - \eta_y \nabla_{\hat{\boldsymbol{\psi}}} \sum_{i \in \mathscr{B}} w_i^y \ell^y \left( \hat{y}_i, y_i \right)$
    end
end

return $\{ \hat{\boldsymbol{\phi}}^c, \hat{\boldsymbol{\phi}}^z, \hat{\boldsymbol{\psi}} \}$
```

5.3 EXPERIMENTAL SETUP

In this section, we describe the experimental setup employed to gauge the predictive performance and interpretability of our models. Our goal is (i) to present a proof of concept for the introduced extensions of the CBMs on simple benchmarks and (ii) apply our methods to a real-world medical image analysis problem, specifically, the prediction of diagnosis, management, and severity from multiview ultrasound scans of pediatric patients with suspected appendicitis. The following subsections introduce synthetic benchmarks and the US imaging dataset. We then motivate and explain baselines, ablation experiments, and evaluation metrics used for model comparison.

5.3.1 *Benchmarking Datasets*

As an initial feasibility study, we experiment with (semi-)synthetic datasets comprising multiview features accompanied by concept and target labels. Similar to Chapter 4, we construct a **synthetic** tabular nonlinear dataset, based on the concept bottleneck model. Its generative process includes (i) sampling the design matrix, (ii) mapping features to concepts, and (iii) utilising these concepts to construct labels. Multiple "views" are generated by retrieving non-overlapping feature subsets from the design matrix. In contrast to the natural image benchmarks for concept-based classification considered in the prior CBM literature, e. g. the Caltech-UCSD Birds-200-2011 [22], [330], our synthetic dataset has per-data-point concept labels instead of assuming class-wide values. Appendix C.1 describes the detailed procedure utilised to generate this synthetic dataset.

In addition to the tabular data, we construct a semi-synthetic attribute-based natural image dataset from *Animals with Attributes* 2 (AwA) [112], [331]. The original AwA consists of 37322 images of 50 animal classes with 85 binary concept variables. The attribute labels are shared across all instances for each class. To investigate the multiview learning scenario, we extend AwA by randomly cropping four patches from each image to produce multiple "views". Appendix C.2 contains a few example images. The resulting **multiview animals with attributes** (MVAwA) dataset has concepts that are partially observable from individual views and assumes no systematic ordering among the patches. Furthermore, similar to the synthetic benchmark, for simplicity, we generate the same number of views for each data point.

5.3.2 *Pediatric Appendicitis Dataset*

Recall that the primary application of interest in the current chapter is ultrasound imaging in pediatric appendicitis (Section 5.1.3). To this end, we study a dataset from a cohort of 579 children and adolescents (0–18 years old) admitted as inpatients to the Department of Pediatric Surgery and Pediatric Orthopedics at the tertiary Children's Hospital St. Hedwig in Regensburg, Germany, between January 1, 2016, and December 31, 2021, with suspected appendicitis. Expanding on the previous analysis by Marcinkevics *et al.* [320], this dataset was acquired and published as part of the current research project [292], [332].

We have collected retrospective data via the hospital's database, including potentially multiple abdominal B-mode ultrasound images for each patient (totalling 1709 images). The number of views per subject ranges from 1 to 15. Usually, images depict various regions of interest, such as the abdomen's right lower quadrant (RLQ), appendix, intestines, lymph nodes, and reproductive organs. Figure 5.3 contains an example of US images belonging to a single patient before and after preprocessing. For each subject, we retrieve ultrasound from admission and initial clinical course with findings related to variables reported in Table C.1 (Appendix C.3).

In addition, we consider information encompassing laboratory tests, physical examination results, and conventional clinical scores, such as Alvarado (AS) and pediatric appendicitis (PAS) scores [333]–[335], widely utilised by pediatricians and pediatric surgeons for the risk stratification of children and adolescents with abdominal pain [336]. Last but not least,

(a)

	D: ***appendicitis***		
S **M**	*compl.*	*uncompl.*	**Total**
surg.	97	135	232
cons.	0	151	151
Total	97	286	383

(b)

	D: ***no appendicitis***		
S **M**	*compl.*	*uncompl.*	**Total**
surg.	0	2	2
cons.	0	194	194
Total	0	196	196

TABLE 5.1: The contingency table of the pediatric appendicitis dataset w.r.t. the management (M; *surgical* vs. *conservative*) by severity (S; *complicated* vs. *uncomplicated*) stratified by the diagnosis (D; *appendicitis* vs. *no appendicitis*).

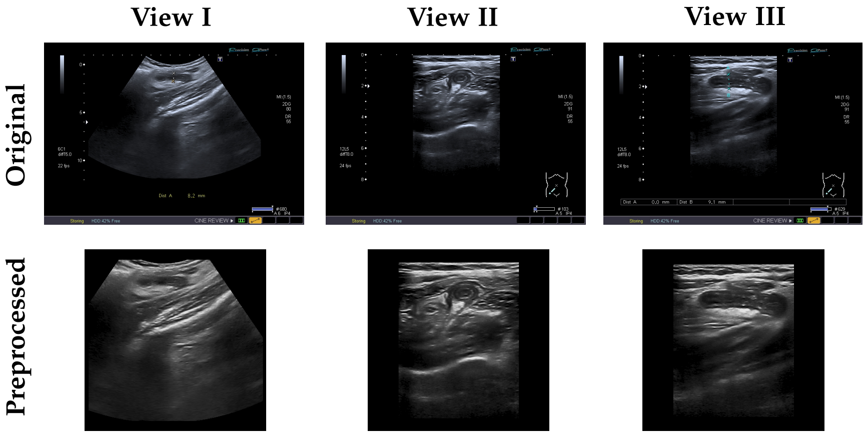


Figure 5.3: An example of multiple US images acquired from a single patient from the pediatric appendicitis dataset. Views I and II correspond to longitudinal and transverse sections of the appendix, respectively, and view III depicts the reaction in the tissue surrounding the appendix. Original images (*top*) contain graphical interface elements and expert-made markers, whereas preprocessed images (*bottom*) have been inpainted, cropped, and padded.

we focus on the expert-produced ultrasonographic findings represented by categorically-valued features. Nine of these were chosen as high-level concepts relevant to decision support. Meaning and comprehensive description of these variables are reported in Table C.1. Throughout this chapter, for brevity, we will denote the concepts by c_1 to c_9.

Each subject is labelled w.r.t. three target variables: (i) diagnosis (*appendicitis* vs. *no appendicitis*), (ii) management (*surgical* vs. *conservative*), and (iii) severity (*complicated* vs. *uncomplicated or no appendicitis*). The frequencies of different label categories and their combinations are shown in Table 5.1. The diagnosis was confirmed histologically in the patients who underwent appendectomy. Subjects treated conservatively are labelled as having appendicitis if their appendix diameter was at least 6 mm and either AS or PAS ≥ 4. Note that the labelling criterion above is only a proxy for the ground-truth disease status, with AS and PAS helping exclude children with no appendicitis [335]. Moreover, the addition of the US information on the enlarged appendix has been shown to increase the positive predictive value [336], [337]. The management label reflects the decision made by a senior pediatric surgeon based on clinical, laboratory, and US data. For the severity, complicated appendicitis includes cases with abscess formation, gangrene, or perforation.

Before model development and evaluation, preprocessing was performed on B-mode ultrasound images to eliminate undesired variability. The study being retrospective, ultrasonograms are as per clinical routine and, therefore, contain graphical user interface elements, markers, distance measurements, and other annotations (Figure 5.3). Consequently, we employ a generative inpainting model, DeepFill [338], to mask and fill such objects. Subsequently, images are resized to 400×400 pixels using zero padding when needed. Finally, contrast-limited histogram equalisation (CLAHE) is applied, and pixel intensities are normalised to the range of 0 and 1. During training, we leverage extensive on-the-fly augmentations to avoid overfitting.

5.3.3 *Baselines and Ablations*

Next to the proposed (SS)MVCBM (Section 5.2), we consider several baselines and ablations. Across all datasets, we apply single-view neural-network-based classifiers. In particular, we train MLPs on tabular data and fine-tune ResNet-18 [339] on images. As an *interpretable* single-view baseline, we employ vanilla CBMs. To ensure a fair comparison between CBMs and (SS)MVCBMs, we utilise identical architectures for individual modules (Appendix C.4). As a black-box multiview baseline, we employ a neural network with the same architecture as that of the MVCBM but trained without concept supervision in the bottleneck layer. We refer to it as the multiview bottleneck (MVBM).

When appropriate, we compare two ways of aggregating per-view representations: averaging and LSTM (Equation 5.3b). Moreover, we consider sequential (Equations 5.4–5.5) and joint (Equation 5.6) training procedures. Finally, on the pediatric appendicitis dataset, we additionally evaluate another baseline—a random forest (RF) [4] trained on radiomic features [269]. This approach is similar to the baseline from Chapter 4 (Section 4.3.2) applied to CT data. Herein, radiomics features are extracted from every US image and averaged across the views for each subject. The performance of this classifier is further improved by ANOVA F-value-based feature selection performed using nested CV.

5.3.4 *Evaluation Metrics*

As the intended use-case scenario for our model is clinical *decision support*, we assess the performance w.r.t. concept and label prediction using AUROC and AUPR. Notably, for pediatric appendicitis, different metrics may be relevant depending on the target variable, e. g. a low false negative rate may be critical for diagnosis and severity. In contrast, a low false positive rate may be desirable for management to avert negative appendectomies [340]. Thus, we do not commit to a single classification threshold. For the pediatric appendicitis experiments, we also report Brier scores to gauge calibration. For concept-based models, in addition to the conventional predictive performance, we assess the effectiveness of interventions (Equation 5.2) by visualising changes in the target AUROC and AUPR across varying percentages of concept variables intervened on. For simplicity, the variables to be edited are chosen at random.

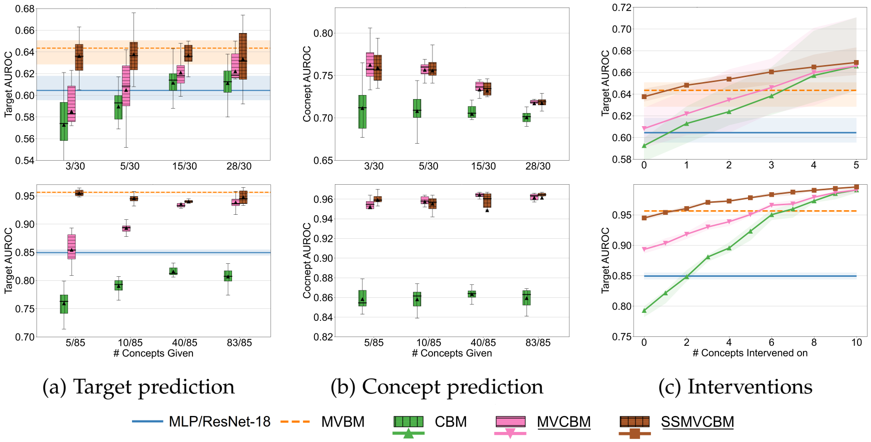
5.4 RESULTS

We now turn to the experimentation results, describing our proof-of-concept findings on the (semi-)synthetic benchmarks followed by a comprehensive analysis on the pediatric appendicitis data. As outlined in the previous subsection, our goal is to (i) demonstrate that, in principle, our models can tackle multiview data under incomplete concept sets and (ii) explore a practical application to medical image analysis.

5.4.1 *Proof of Concept on Synthetic Data*

We first benchmark our methods on tabular synthetic nonlinear data (Appendix C.1) and multiview animals with attributes (Appendix C.2). In both experiments, we train models with varying numbers of concepts observed to emulate the incomplete concept set scenario (Figure 5.2), where some attributes are systematically missing. Throughout this subsection, we focus on the AUROC for predictive performance evaluation, as we observed similar results w.r.t. AUPR, which we omit in the interest of space.

Figure 5.4 contains the summary of the results, with the top row corresponding to the synthetic data. Expectedly, black-box and concept-based *multiview* approaches are consistently more accurate than their single-view counterparts at target (Figure 5.4a) and concept prediction (Figure 5.4b).



(a) Target prediction (b) Concept prediction (c) Interventions

FIGURE 5.4: Results on the synthetic (*top*) and MVAwA (*bottom*) datasets for the proposed **multiview concept bottleneck (MVCBM)** and **semi-supervised multiview concept bottleneck (SSMVCBM)** models alongside several baselines. (a) One-vs-all AUROCs for predicting the target under the varying number of observed concepts. Note that MLP, ResNet-18, and MVBM do not rely on concepts; their AUROCs are shown as horizontal lines for reference. (b) Averaged AUROCs for concept prediction. (c) AUROCs for target prediction after intervening on the varying number of concepts. Interventions are performed under an incomplete concept set: 5/30 and 10/85 observed concepts for the synthetic and MVAwA datasets. The performance of non-intervenable baselines is shown by horizontal lines.

Thus, a multiview black-box model without concept supervision (MVBM) performs considerably better than an MLP trained on a single view. Likewise, an MVCBM outperforms a simple CBM. Notably, the target prediction performance for CBM and MVCBM increases with the number of concepts observed, as shown in Figure 5.4a. When a near-complete concept set is provided, the performance of the multiview CBM is close to that of the multiview black box. The SSMVCBM performs well even when few concepts are known and is close to the black-box baseline across most settings.

For concept prediction (Figure 5.4b), MVCBM and SSMVCBM attain comparable performance with higher AUROCs than the single-view version. As expected, the semi-supervised model predicts the concepts equally well compared to the MVCBM. Thus, representation learning has no effect on the concept prediction. Lastly, we observe from Figure 5.4c that, similarly to the classical CBM, both multiview variants allow for effective interventions, i. e.

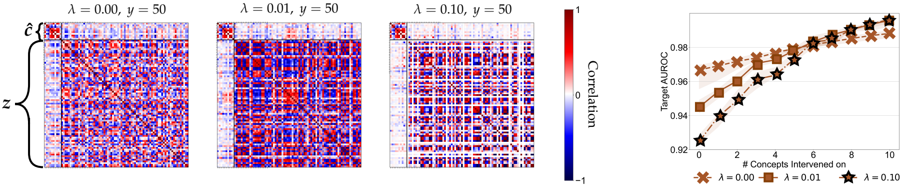
their predictive performance improves when replacing predicted concepts with the ground truth at test time.

For the MVAwA dataset, we observe analogous results summarised in the bottom panels of Figure 5.4. In particular, (i) multiview techniques perform superior to single-view approaches, (ii) under a complete concept set, MVCBM is comparable to the black box, and (iii) MVCBM and SSMVCBM can be effectively intervened on.

5.4.2 *The Role of Adversarial Regulariser*

In addition to the results above, we conduct an ablation study on the SSMVCBM to investigate the effect of adversarial regularisation (Equation 5.7). As explained in Section 5.2.3, adversarial regulariser helps decorrelate learned representations $\boldsymbol{z}$ and predicted concept values $\hat{\boldsymbol{c}}$. Below, we explore the role of this regulariser by training models under varying hyperparameter (λ) values on the MVAwA dataset.

To this end, we assess the correlation among predicted concept values and representations and intervention effectiveness, as shown in Figure 5.5. Expectedly, stronger regularisation ($\lambda > 0.00$) hurts the target prediction performance (Figure 5.5b) but allows learning representations de-correlated from the concepts (Figure 5.5a). Nevertheless, even in the absence of adversarial regularisation ($\lambda = 0.00$), $\hat{\boldsymbol{c}}$ and $\boldsymbol{z}$ are already merely *weakly* correlated. Importantly, regularised models demonstrate a steeper increase



(a) (b)

FIGURE 5.5: Results of the ablation study on the effect of the adversarial regularisation in the semi-supervised multiview concept bottleneck model (SSMVCBM). SSMVCBM models were trained on the MVAwA dataset under varying regularisation parameter $\lambda = 0.00, 0.01, 0.10$. (a) Conditional correlation among the predicted concepts ($\hat{\boldsymbol{c}}$) and representations ($\boldsymbol{z}$) for class $y = 50$. (b) Intervention results for varying regularisation strength.

in predictive performance during interventions, eventually attaining a higher AUROC than the unregularised SSMVCBM. In summary, adversarial regularisation helps learning representations disentangled from concept variables and improves the effectiveness of interventions, albeit at the price of performance without interventions.

5.4.3 *Application to Pediatric Appendicitis*

Multiview concept bottleneck models are readily applicable to medical imaging datasets, which often include multiple views or heterogeneous data types. In this subsection, we turn to a more realistic scenario and explore the application of multiview CBMs to pediatric appendicitis data (Section 5.3.2).

To begin with, we evaluate the ability of all concept-based models to predict high-level appendix ultrasound features (Table C.1) from (multiple) abdominal US images. Table 5.2 reports test-set AUROCs and AUPRs achieved by the different variants of the concept bottleneck. In addition to different designs, we investigate the effect of the optimisation procedure, sequential vs. joint (Equations 5.4–5.6), and view-specific feature fusion, averaging vs. LSTM. Note that Tables 5.2a and 5.2b focus on the diagnosis as the target variable. We observe similar results for the management and severity reported in Tables C.3 and C.4 (Appendix C.5). We attribute minor discrepancies across the three classification problems to the differences in the weights assigned to data points in the cost-sensitive loss function (Equations 5.4–5.6 and 5.7) and the choice of hyperparameter values.

Across all target variables, most concepts could be predicted by at least one of the models significantly better than by a fair coin flip (one-sample two-sided t-test p-value < 0.05, adjusted using the Benjamini–Yekutieli procedure [341] with the FDR of $q = 0.05$). Surprisingly, some of the variables with relatively few cases present in the dataset could be captured by some models, e. g. *coprostasis* (c_8) and *meteorism* (c_9) by the LSTM-based variants of MVCBM and SSMVCBM. By contrast, the *thickening of the bowel wall* (c_7) was particularly challenging to model, likely due to its low prevalence and the lack of predictive power in the downstream classification task.

Predictably, sequentially optimised models (seq) are more performant at the concept prediction than the ones optimised jointly (joint), in agreement with the findings reported in the literature [22]. Similar to the results from Figure 5.4b, models aggregating *multiple* views tend to have higher AUROCs and AUPRs. In addition, LSTM-based aggregation consistently

(a)

Model	Concept AUROC								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
CBM-seq	0.52±0.04	0.47±0.04	0.60±0.07*	0.56±0.08	0.63±0.05*	0.57±0.05*	0.45±0.08	0.48±0.08	0.39±0.07
CBM-joint	0.50±0.05	0.47±0.03	0.57±0.05*	0.54±0.06	0.64±0.04*	0.59±0.05*	0.39±0.06	0.57±0.12	0.38±0.09
MVCBM-seq-avg	0.61±0.05*	0.49±0.05	0.66±0.08*	0.60±0.08*	0.51±0.08	0.66±0.08*	*0.50±0.04*	0.47±0.12	0.55±0.07
MVCBM-seq-LSTM	*0.83±0.03**	*0.59±0.03**	0.62±0.04*	**0.71±0.04***	0.65±0.04*	0.67±0.07*	0.49±0.07	**0.68±0.10***	*0.73±0.06**
MVCBM-joint-avg	0.55±0.10	0.47±0.07	**0.73±0.07***	0.63±0.07*	0.61±0.06*	0.63±0.07*	0.48±0.06	0.45±0.13	0.54±0.11
MVCBM-joint-LSTM	**0.85±0.03***	0.55±0.04*	0.58±0.04*	*0.70±0.03**	**0.75±0.02***	0.55±0.09	0.45±0.12	0.68±0.17	**0.77±0.03***
SSMVCBM-avg	0.62±0.05*	**0.60±0.05***	*0.72±0.05**	0.67±0.05*	0.54±0.05	*0.68±0.08**	**0.53±0.11**	0.43±0.08	0.47±0.07
SSMVCBM-LSTM	**0.85±0.04***	0.58±0.06*	0.66±0.05*	**0.71±0.06***	*0.67±0.04**	**0.69±0.06***	0.45±0.09	*0.66±0.11**	*0.73±0.05**

(b)

Model	Concept AUPR								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Random	0.72	0.49	0.19	0.23	0.51	0.26	0.16	0.13	0.14
CBM-seq	0.71±0.03	0.53±0.03*	0.29±0.06*	0.26±0.05	0.64±0.05*	0.38±0.06*	0.15±0.03	0.12±0.02	0.11±0.02
CBM-joint	0.73±0.05	0.49±0.04	0.30±0.06*	0.30±0.08	0.64±0.05*	0.38±0.09*	0.15±0.05	0.19±0.08	0.11±0.02
MVCBM-seq-avg	0.79±0.04*	0.53±0.06	*0.34±0.10**	0.35±0.10*	0.53±0.07	*0.41±0.07**	0.17±0.04	0.14±0.04	0.25±0.12
MVCBM-seq-LSTM	0.92±0.02*	*0.59±0.04**	0.32±0.05	**0.38±0.04***	*0.67±0.04**	**0.42±0.10***	0.15±0.02	*0.21±0.08*	**0.40±0.11***
MVCBM-joint-avg	0.75±0.08	0.48±0.06	**0.38±0.09***	0.30±0.06	0.58±0.05*	0.39±0.08*	**0.21±0.08**	0.15±0.08	0.16±0.05
MVCBM-joint-LSTM	**0.94±0.01***	0.50±0.05	0.26±0.08	*0.37±0.07**	**0.74±0.04***	0.32±0.09	0.16±0.08	**0.31±0.20**	0.28±0.07*
SSMVCBM-avg	0.79±0.04*	0.58±0.03*	**0.38±0.05***	0.34±0.04*	0.54±0.06	**0.42±0.08***	*0.20±0.06*	0.12±0.04	0.17±0.07
SSMVCBM-LSTM	*0.93±0.03**	**0.60±0.06***	0.31±0.06*	**0.38±0.06***	*0.67±0.04**	0.39±0.06*	0.19±0.06	0.19±0.07	*0.30±0.09**

TABLE 5.2: Concept prediction performance on the pediatric appendicitis dataset with the *diagnosis* as the target variable. (a) AUROCs and (b) AUPRs are reported as averages and standard deviations across ten independent initialisations. Herein, "seq" and "joint" denote sequential and joint optimisation, respectively, whereas "avg" and "LSTM" stand for the averaging- and LSTM-based fusion. Averages significantly greater than the expected performance of a fair coin flip (random) are marked by "*". **Bold** indicates the best result, and *italics* indicates the second best. The meaning of the concept variables: c_1, visibility of the appendix; c_2, free intraperitoneal fluid; c_3, appendix layer structure; c_4, target sign; c_5, surrounding tissue reaction; c_6, pathological lymph nodes; c_7, thickening of the bowel wall; c_8, coprostasis; c_9, meteorism.

Model	**Diagnosis**			**Management**			**Severity**		
	AUROC	**AUPR**	**Brier**	**AUROC**	**AUPR**	**Brier**	**AUROC**	**AUPR**	**Brier**
Random	0.50	0.75	0.25	0.50	0.47	0.25	0.50	0.23	0.25
Radiomics + RF	0.64±0.02	0.82±0.01	0.22±0.00	0.65±0.01	0.60±0.02	*0.24±0.00*	*0.77±0.02*	*0.58±0.04*	**0.15±0.00**
ResNet-18	0.70±0.07	0.88±0.04	0.25±0.08	0.69±0.07	*0.71±0.08*	0.27±0.05	0.73±0.10	0.52±0.10	0.18±0.04
CBM-seq	0.64±0.06	0.84±0.04	0.22±0.02	0.68±0.05	0.68±0.05	**0.23±0.02**	0.66±0.06	0.41±0.08	0.23±0.04
CBM-joint	0.62±0.04	0.83±0.04	0.24±0.02	0.66±0.06	0.68±0.04	**0.23±0.02**	0.68±0.06	0.44±0.08	0.23±0.02
MVBM-avg	*0.76±0.05*	0.89±0.04	0.22±0.03	0.71±0.04	0.69±0.04	0.24±0.02	0.71±0.12	0.59±0.11	0.20±0.05
MVBM-LSTM	*0.76±0.04*	*0.91±0.02*	0.23±0.02	0.67±0.04	0.61±0.04	**0.23±0.02**	0.74±0.13	*0.58±0.12*	0.22±0.07
MVCBM-seq-avg	0.67±0.05	0.85±0.05	0.23±0.02	0.58±0.05	0.62±0.06	0.26±0.02	0.75±0.07	0.56±0.12	0.23±0.04
MVCBM-seq-LSTM	0.73±0.03	0.89±0.01	0.24±0.04	0.57±0.03	0.53±0.04	0.26±0.01	0.70±0.11	0.48±0.16	0.21±0.03
MVCBM-joint-avg	0.66±0.09	0.84±0.06	0.24±0.06	0.69±0.06	0.66±0.11	**0.23±0.02**	0.70±0.06	0.53±0.11	0.24±0.02
MVCBM-joint-LSTM	0.72±0.02	0.88±0.02	0.22±0.01	0.57±0.05	0.50±0.04	0.26±0.01	0.65±0.07	0.37±0.10	0.24±0.02
SSMVCBM-avg	**0.80±0.03**	**0.92±0.02**	*0.20±0.03*	**0.72±0.05**	**0.72±0.04**	0.27±0.05	0.73±0.07	0.57±0.09	*0.17±0.02*
SSMVCBM-LSTM	**0.80±0.06**	**0.92±0.04**	**0.19±0.04**	0.70±0.03	0.67±0.06	0.27±0.04	**0.78±0.05**	*0.58±0.10*	0.21±0.10

TABLE 5.3: Target prediction results for the diagnosis, management, and severity.

and noticeably outperforms simple averaging (avg), especially for predicting the visibility of the appendix—one of the most important diagnostic concepts [320]. This could be associated with loose spatiotemporal ordering among US images acquired for each subject. Finally, semi-supervised bottlenecks have concept prediction performance comparable to the sequentially optimised MVCBMs.

As explained before, our end goal is the prediction of the (i) diagnosis, (ii) management, and (iii) severity among suspected appendicitis patients. In Table 5.3, we explore the predictive performance for these three target variables. With respect to AUROC and AUPR, all models are able to predict all targets better than the naïve baseline. Among concept-based approaches, multiview models offer a consistent improvement over the vanilla CBM for diagnosis and severity. Moreover, the best-performing concept-based classifiers often achieve AUROCs and AUPRs comparable to those of the multiview black box.

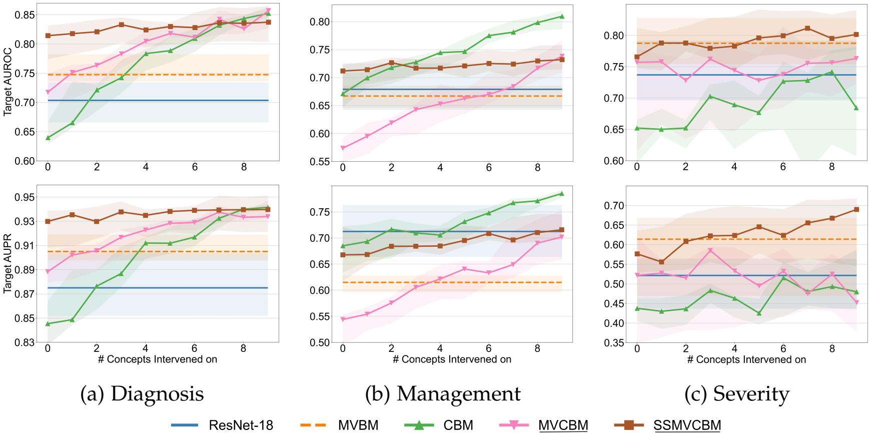
On average, MVCBMs with the LSTM-based fusion outperform averaging-based approaches for diagnosis. However, the opposite is true for management. For diagnosis and management prediction, we also observe that neural-network-based methods outperform RFs trained on radiomics features. The latter result is not surprising, given that we do not utilise manually segmented regions of interest for radiomics feature extraction. Lastly, across all targets, the semi-supervised extension of the MVCBM achieves higher AUROCs and AUPRs or is comparable to the approaches that purely rely on the concepts.

Brier scores partially agree with AUROCs and AUPRs. However, they feature less variability across model classes. For all target variables, most scores are ≥ 0.20. Combined with the reported AUROCs and AUPRs, the latter finding indicates that the probabilistic predictions of the models considered could benefit from calibration, which could help produce more interpretable probabilistic outputs.

To summarise, concept-based classification on multiview US data is encouragingly effective at predicting the diagnosis. For management, aggregating multiple US images offers no improvement over simple single-view classification. We attribute this to the *diagnostic* nature of the chosen concepts and their limited predictive power for the treatment assignment. Likewise, accurately predicting appendicitis severity is challenging, likely due to the low prevalence of complicated appendicitis cases in the current dataset (Table 5.1). Last but not least, in all tasks, the proposed SSMVCBM successfully mitigates the poorer discriminative performance of concept-based approaches by learning representations complementary to the probably incomplete concept set.

Similar to the proof-of-concept experiments (Figure 5.4c), we intervene on the bottleneck layers of the CBM, MVCBM, and SSMVCBM trained on the pediatric appendicitis data. Figure 5.6 contains a summary of our findings. In these experiments, we utilise LSTM-based fusion since it led to better concept prediction performance (Table 5.2). In particular, Figure 5.6a shows intervention results for the diagnosis, and Figures 5.6b and 5.6c correspond to the other two target variables. Analogous to Figure 5.4c, lines show changes in median AUROC and AUPR when intervening on randomly chosen concept subsets of varying sizes.

For the diagnosis (Figure 5.6a), interventions affect the behaviour of the models as in the experiments on the synthetic and natural image datasets. Specifically, AUROC and AUPR increase steadily with the number of concepts intervened on: across all models, the maximum median AUROC and AUPR attained are approx. 0.85 and 0.94, respectively. As the best-performing model (Table 5.3), SSMVCBM demonstrates only a slight increase in median predictive performance after intervening on the full concept set.



(a) Diagnosis (b) Management (c) Severity

FIGURE 5.6: Intervention results for the (a) diagnosis, (b) management, and (c) severity w.r.t. AUROC (*top*) and AUPR (*bottom*). The performance of non-intervenable ResNet-18 and MVBM baselines is shown as horizontal lines.

Likewise, for management (Figure 5.6b), we observe an increase in AUROC and AUPR. However, a single-view CBM performs well and overtakes both multiview models after interventions. Lastly, interventions yield no visible performance improvement for severity (Figure 5.6c), possibly due to considerable variance across initialisations and randomly sampled concept subsets.

In addition to the experimental findings, another output of this chapter is an online decision support tool developed based on the introduced methods. As a step towards informing clinicians and other interested parties about ML-based decision support, we make this tool publicly available at `https://papt.inf.ethz.ch/mvcbm`. Figure C.2 in Appendix C.6 contains a summary with an illustrative use-case example. The tool utilises the multiview CBM model (Figure 5.1) for predicting the *diagnosis*, as we observed the most promising results for this target variable and model configuration (Table 5.3).

Via the tool's simple web interface, the user may upload several ultrasonography images acquired from the same patient. Image preprocessing, as described in Section 5.3.2, may be optionally executed. In addition to prediction, the tool allows intervening on the concept values by editing corresponding sigmoid activations.

5.5 DISCUSSION

Below, we reflect on the contributions of the current chapter and its relation to the general topic of the thesis. We emphasise the relevance of the presented methods in comparison to the related research and summarise valuable empirical contributions. We conclude this section with a discussion of limitations and open questions for future work.

In complement to Chapters 3 and 4, we have explored another class of interpretable models that, in contrast to raw features or prototypes, relies on high-level attributes to explain complex input-output relationships [22], [111], [112]. Specifically, we have concentrated on the class of concept bottleneck models [22] (Sections 2.3.4 and 5.1.1), addressing their practical limitations in the scope of the biomedical application domain.

Motivated by challenges arising in medical image analysis, we generalised conventional CBMs to datasets with *multiple* views and modalities. To this end, we have proposed a practical architecture (Section 5.2.1, Figure 5.1) building on the hybrid fusion approach to multimodal learning [33]. Our model effectively handles (i) varying numbers of views per data point, (ii) partial observability of concepts from individual views (Section 5.1), and (iii) exploits spatial and temporal ordering among images.

Another limitation of CBMs explored in this chapter is the implicit assumption of a sufficient, or complete, concept set [135]. To this end, we have investigated the setting where concept variables are systematically missing and, hence, the observed attributes do not comprehensively capture the relationship between the covariates and the target. To tackle this issue, we have introduced another model design (Section 5.2.3) combining concept prediction with representation learning. Additionally, to disentangle representations and concept predictions and improve intervention effectiveness, we have utilised adversarial regularisation (Equation 5.7), adapting the model's training procedure accordingly (Algorithm 2).

Despite model design adjustments, our multiview concept bottleneck and its semi-supervised variant retain the interpretability of conventional CBMs. Firstly, our models follow the same structure of successively predicting concepts from covariates and the target from the concepts. Secondly, adversarial regularisation allows de-correlating concepts from opaque representations, thus improving predictive performance without unwanted redundancies. Furthermore, as demonstrated empirically (Figures 5.4c and 5.5b and Figure 5.6c), interventions on predicted concepts are still effective and allow increasing test-time predictive performance through human–model interaction.

Many previous works have investigated related limitations of the CBMs and explored alternative model designs. For instance, Sawada and Nakamura [342] combine CBMs with self-explaining neural networks (Section 2.3.4) and incorporate additional unsupervised concepts. However, they do not explore the disentanglement of the given and learnt concepts and do not provide results on the effectiveness of interventions.

Yüksekgönül, Wang, and Zou [343] propose training a concept bottleneck *post hoc* on the representations from a pretrained backbone network. Additionally, they utilise residual fitting to compensate for an incomplete concept set by adding a residual term to the bottleneck's final prediction. In contrast, this chapter studies the *ante hoc* modelling scenario and provides a solution technically different from residual modelling, in line with the work by Sawada and Nakamura [342].

Another related line of work tackles unobserved concepts and leakage [344] by employing representation learning and generative modelling. Despite facilitating probabilistic modelling and conditional generation, e. g. as in Chapter 4, generative approaches can be data-hungry and, thus, ineffective on small and high-dimensional datasets, such as pediatric appendicitis (Section 5.3.2).

Last but not least, the concurrent method by Havasi, Parbhoo, and Doshi-Velez [345] extends standard CBMs by including a side channel similar to the representation learning branch of the SSMVCBM and capturing autoregressive relationships among the concepts. In contrast, we additionally tackle the multiview learning scenario (Section 5.1.2) and medical imaging applications, whereas Havasi, Parbhoo, and Doshi-Velez [345] consider ICU time series data.

From the multiview and multimodal learning perspective, we capitalise on prior research on aggregating representations across views or modalities via averaging [325] and sequential modelling [326]. Nevertheless, alternative fusion approaches exist, e. g. products and mixtures of experts [243], [289]. To the best of our knowledge, this chapter is among the few efforts at designing *interpretable* multiview models for concept-based classification.

Beyond comprehensive evaluation conducted on the synthetic and natural image datasets (Sections 5.4.1–5.4.2), the primary empirical contribution of this chapter is the investigation of ultrasonography-based ML models for pediatric patients with suspected appendicitis. Although appendicitis is a common condition in the pediatric population, diagnosing it and choosing the best therapeutic option is challenging. Early differentiation between simple and complicated necrotising appendicitis is crucial for effective management and prognosis [303], [315], [346]. Thus, ML-based decision support tools, such as those explored in this chapter, may increase diagnostic accuracy and prove pivotal in improving treatment outcomes. Our empirical findings (Section 5.4.3) promisingly suggest that the direct interpretation of US images by ML models is feasible. Such predictive models may assist physicians in interpreting US images and enable comparison of the results with the newly conducted US exams to characterise the progress or resolution of the inflammation.

Most of the prior work on using ML for appendicitis has focused on tabular datasets with handcrafted features [313]–[318], [320]–[322] or more invasive imaging modalities, such as CT [319]. In contrast, we take the first steps towards the computer-aided diagnosis of appendicitis based on abdominal *ultrasound*, a noninvasive, accessible, and cheap technique. Additionally, we have publicised an anonymised version of the dataset utilised for model development to facilitate future replication of our findings and method comparison [332]. Lastly, for demonstratory and educational purposes, we have deployed our MVCBM for predicting the diagnosis as an easy- and free-to-use web tool (Appendix C.6).

In our experiments (Section 5.4), we have demonstrated the benefits of the multiview and semi-supervised concept-based approach on the synthetic tabular, natural, and medical image data. Our findings show that multiview concept bottlenecks and their semi-supervised variant generally outperform vanilla CBMs in concept and target prediction (Figure 5.4).

For the diagnosis prediction on the pediatric appendicitis dataset, MVCBMs attain performance comparable to black-box multiview classifiers (Table 5.3) while facilitating the interpretation of and interventions on predictions via clinically relevant attributes (Table C.1). For management and severity, the results are somewhat less conclusive, featuring little difference across single- and multiview models. We attribute the latter findings to the limited predictive power of the ultrasonographic features for these target variables [320], the diagnostic nature of the chosen concepts, and the moderate size of the cohort (Table 5.1). For instance, previous literature [320] has shown that the most important predictor of the treatment assignment is peritonitis/abdominal guarding assessed during a clinical examination. Among the US findings, most other predictively useful attributes can be identified based on the right lower quadrant image alone. Therefore, we hypothesise that the additional views, e. g. depicting pathological lymph nodes or meteorism, are not as helpful for the management classification. This observation might explain the relatively worse performance of the multiview approaches for this target variable.

5.5.1 *Limitations*

Despite consistent improvements in empirical performance, semi-supervised multiview concept bottlenecks have a few limitations. Although the representations learnt are de-correlated from the concept variables, their semantic meaning remains opaque and further investigation is necessary to make this model class fully interpretable. It is also unclear if these representations can recover the ground-truth unobserved concepts (Figure 5.2) up to some reasonable transforms.

As evidenced by the experimental findings, models' predictions are not always well-calibrated (Table 5.3). Potentially, a probabilistic treatment of the concept and target variables may constitute a more principled approach to uncertainty estimation. A probabilistic model, akin to the one from the previous chapter, would improve the interpretability of predictions and, for example, allow for selective classification [347] and uncertainty-based interventions [348].

In regard to the pediatric appendicitis dataset, our study's design and experimental setup likewise have weaknesses. The dataset was acquired from a moderately sized and relatively homogeneous patient cohort recruited from a single clinical centre over a short time period. Hence, external validation on the data from other US devices, clinical centres, and countries is necessary to test the generalisability of our models.

Another limitation is the lack of histologically confirmed diagnoses among conservatively treated patients. Thus, the model validation and comparison results must be interpreted cautiously since we do not have access to the true disease status of many subjects. Likewise, our current US image preprocessing pipeline is imperfect: we crop and resize images (Figure 5.3), making it impossible to detect the appendix diameter, a relevant sonographic sign of appendicitis [346]. Finally, the current analysis does not incorporate data from multiple raters or physicians' uncertainty.

5.5.2 *Future Work*

We envision many enhancements and extensions to the multiview concept bottlenecks and their application to pediatric appendicitis and, generally, medical imaging. Most of these research directions are closely linked to the limitations discussed before.

In particular, it would be interesting to explore model design alterations beyond the scope of the current chapter. As adaptive fusion using LSTMs has proven effective, we would like to investigate alternative learnable functions, including those invariant to the view order. As mentioned above, the representation learning branch of the SSMVCBM is a black box, and, thus, introducing additional regularisation to, for instance, de-correlate individual dimensions of representations [328] might aid the interpretability. Furthermore, principled uncertainty quantification can further enhance the practical utility of our models. For example, proposed architectures can be combined with the modules from stochastic segmentation networks [349] or probabilistic concept bottlenecks [350].

For the pediatric appendicitis application, comprehensive predictive models should incorporate clinical and laboratory parameters and consider the presence of other conditions, such as COVID-19. Additionally, more refined definitions for the target variables could provide further insights, e. g. differentiating between subacute and acute appendicitis cases for the diagnosis and predicting the risk of a secondary appendectomy for the management. Lastly, adjustments in the model architecture and the acquisition of a larger training dataset can facilitate the incorporation of the colour Doppler images, potentially making the prediction of the disease severity progression more accurate.

5.6 SUMMARY

The current chapter considered concept-based classification models that link the high-dimensional feature space and response via high-level attributes, or concepts, interpretable to human experts. While the literature features a plethora of concept-based model designs and outlines the limitations of this class, many enhancements are necessary to translate these methods into the "wild" of the biomedical application domain.

Inspired by the characteristics of medical imaging datasets, we (i) proposed model designs allowing for concept-based classification on *multiview* data and (ii) considered the problems where the set of concepts given at training time is *incomplete*, i. e. some attributes are *systematically* missing. Such scenarios are practically relevant due to the complexity and multimodality of biomedical data, the potential incompleteness of domain knowledge, and the costs associated with measurements. Our models adhere to the original interpretable structure of successively predicting concepts from the covariates and targets from the concepts. Moreover, they enable effective human–model interaction, wherein an expert may edit predicted concepts to "steer" the downstream output.

We provided proof-of-concept experimental results on synthetic tabular and semi-synthetic natural image datasets, demonstrating the utility of our methods. The primary empirical contribution of this chapter is the application of our concept-based classifiers to the prediction of the diagnosis, management, and disease severity in pediatric patients with suspected appendicitis based on abdominal ultrasound images. Developing more standardised criteria for the diagnosis and management of this disease remains an open challenge, with the majority of previous work concentrating on tabular data or more invasive and costly imaging modalities. Our efforts demonstrate that in this context, concept-based classifiers can be successfully leveraged for predictive modelling, decision support, and medical image interpretation.

Part II

POST HOC METHODS

INTERMEZZO: POST HOC METHODS

The previous part of this thesis has treated the design of *ante hoc* interpretable models. Although this perspective is conceptually compelling, for practical reasons, we often have to resort to black boxes, explaining their predictions *post hoc* (Section 2.4). For instance, this may be necessitated by the lack of domain knowledge at the training time or the considerably better predictive performance of a black-box model compared to interpretable alternatives. In the second part of this manuscript, we will focus on the *post hoc* explanation techniques and leverage them to (i) interact with (Chapter 6) and (ii) edit (Chapter 7) neural networks. Thus, Chapters 6 and 7 will address Question 2, raised at the beginning of the thesis. Specifically, Chapter 6 will explore the utility of probing for concept-based interventions akin to those facilitated by concept bottleneck models (Chapter 5). In Chapter 7, we will investigate the use of attribution-based methods (Section 2.4.1) to mitigate biases in neural network classifiers and attain algorithmic fairness. In addition to conventional benchmarking datasets, both chapters will demonstrate the utility of the proposed methods on publicly available chest radiograph data.

6

BEYOND CONCEPT BOTTLENECKS

In Chapter 5, we treated the problem of translating concept-based models to the realities of biomedical data, introducing several practical and empirically effective enhancements. Compelling as they are, concept bottleneck models [22] have an apparent limitation: they require concept knowledge and annotated data at *training time*. In this chapter, we turn to the *post hoc* explanation setting and devise a method to perform *concept-based interventions* on an *already* trained neural network model with a black-box architecture.

Let us recall the pediatric appendicitis dataset from Chapter 5 as a concrete motivating example for the methods to be presented below. In high-throughput medical image examination, we might resort to utilising a black-box predictive model to support medical professionals' decisions. In many cases, designing an interpretable model from scratch may be impossible or impractical. However, a medical professional typically has the knowledge of clinically relevant high-level attributes, and we may be able to acquire a moderately sized annotated and labelled validation set. In such a scenario, the techniques we will introduce allow the professional to interact with the black box via human-understandable concepts, potentially establishing a high-level understanding of the classifier's behaviour and steering its predictions to achieve the desired outcome.

Contributing to the prior efforts at converting black boxes into CBMs [343], [351], we focus on the interventions (Section 5.1.1) and human–model interaction that form a unique aspect distinguishing CBMs from many other interpretable model classes (Section 2.3.4). In particular, we investigate (i) how to perform instance-specific concept-based interventions on black-box models and (ii) how to formally quantify and improve the effectiveness of such interventions. We believe that, in addition to facilitating interaction, interventions fit into the broader family of *post hoc* explanation techniques (Section 2.4), as they enable mechanistic model understanding through *editing* in line with the argument by Arora *et al.* [24], who study model editing "exercises" in the context of evaluating explanations.

In the following sections, we summarise closely related works, adding to the discussion from Section 5.1. We then introduce a procedure for concept-based interventions on black-box neural networks and formalise the notion of intervention effectiveness. In addition, we introduce a procedure for explicitly fine-tuning the models to increase the intervention impact. Subsequently, in the experiments, we explore our methods on the synthetic tabular, natural image, and medical image datasets, benchmarking them against a few common-sense and related techniques. At the end, we provide a general discussion of our findings and contributions, outlining current limitations and potential questions for future research. This chapter's content and text are based on the preprint "Beyond Concept Bottleneck Models: How to Make Black Boxes Intervenable?" [352].

6.1 BACKGROUND

This section introduces preliminaries and the broader context of the literature related to the methods described in this chapter. To begin with, let us formally outline the setting and notation adhered to. Similar to Chapter 5, let $\boldsymbol{x}$, $\boldsymbol{c}$, and y denote the covariates, concepts, and targets, respectively. As in Equation 5.1, a concept bottleneck model [22] is given by $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\psi}}(h_{\boldsymbol{\phi}}(\boldsymbol{x})) = g_{\boldsymbol{\psi}}(\hat{\boldsymbol{c}})$, where $h_{\boldsymbol{\phi}}$ is trained to predict the concepts and $g_{\boldsymbol{\psi}}$ maps the predicted concept values to the target. CBMs are trained on labelled and annotated data $\{(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)\}_i$ using the procedures discussed in Section 5.2.2. Likewise, for a *black-box* model $f_{\boldsymbol{\theta}}$, similar to Definition 2.4.5, we consider a slice $\langle g_{\boldsymbol{\psi}}, h_{\boldsymbol{\phi}} \rangle$, s.t. $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\psi}}(h_{\boldsymbol{\phi}}(\boldsymbol{x}))$. Thus, $\boldsymbol{z} = h_{\boldsymbol{\phi}}(\boldsymbol{x})$ denotes the activations, or representation, in the layer defined by the slice. We assume that the black-box model has been trained end-to-end on labelled data $\{(\boldsymbol{x}_i, y_i)\}_i$, i. e. *without* explicit exposure to the concepts.

6.1.1 *Related Work*

In addition to the literature discussed in Chapter 5, we highlight a few relevant lines of research below. Building on the CAVs and conceptual sensitivity by Kim *et al.* [129] (Section 2.4.3, Definition 2.4.5), Abid, Yuksekgonul, and Zou [353] introduce conceptual counterfactual explanations (CCE) whose goal is to identify concept variables that may induce the change in the label predicted by a neural network classifier.

Most related to the topic of the current chapter are the works on transforming black-box models into a CBM-like structure [343], [351]. Specifically, *post hoc* CBMs by Yüksekgönül, Wang, and Zou [343] project backbone representations into the concept space constructed using CAVs to subsequently predict the target variable. The authors also investigate the utility of an *additive* residual channel to supplement concept-based prediction and recover the black box's predictive performance. Label-free CBMs [351] adopt a similar approach but, instead of the CAV framework, utilise embedding similarities. In addition, both works explore the use of multimodal models, such as CLIP [354], to construct concept labels in the absence of annotated data.

Given the focus of this chapter, another relevant research direction concentrates on interventions and improving their effectiveness in CBMs. For instance, Shin *et al.* [348] conduct a thorough empirical comparison among different intervention procedures for CBMs. Similarly, Sheth *et al.* [355] and Chauhan *et al.* [356] introduce *adaptive* strategies to enhance intervention effectiveness at test time. Lastly, Steinmann *et al.* [357] propose learning to detect mistakes among predicted concept values and correcting these hypothesised errors.

6.2 CONCEPT-BASED INTERVENTIONS ON BLACK-BOX MODELS

Following the notation from Section 6.1, we now introduce a technique for performing concept-based interventions on black-box neural networks. Furthermore, we formalise a measure for the effectiveness of such interventions and propose a procedure to explicitly fine-tune the model for this measure. For all our techniques, we will assume being given a *validation* set $\mathscr{D}_{\text{valid}} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)\}_{i=1}^{N}$ comprising N labelled and annotated data points. Note that, in practice, this dataset can be considerably smaller than the original training set and, as proposed in the literature, concept labels can be, in principle, generated using vision-language models (VLM) [343], [351].

In the current setting, given a black-box model $f_{\boldsymbol{\theta}}$ and a data point $(\boldsymbol{x}, y)$, a human user might desire to influence the prediction $\hat{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ via high-level and understandable concept values $\boldsymbol{c}'$. In the high-throughput medical image examination example discussed at the beginning of this chapter, a medical professional may want to interact with the classifier ($f_{\boldsymbol{\theta}}$) by annotating their findings ($\boldsymbol{c}'$) in the image ($\boldsymbol{x}$). To facilitate such interactions, we propose a simple recipe for concept-based instance-specific interventions summarised schematically in Figure 6.1. Our technique can be applied to any black-box neural network. Intuitively, using given validation data and concept values, this procedure edits the network's representations $\boldsymbol{z} = h_{\boldsymbol{\phi}}(\boldsymbol{x})$ to align them more closely with $\boldsymbol{c}'$ and, thus, affect the downstream prediction. Below, we explain the individual steps of the method.

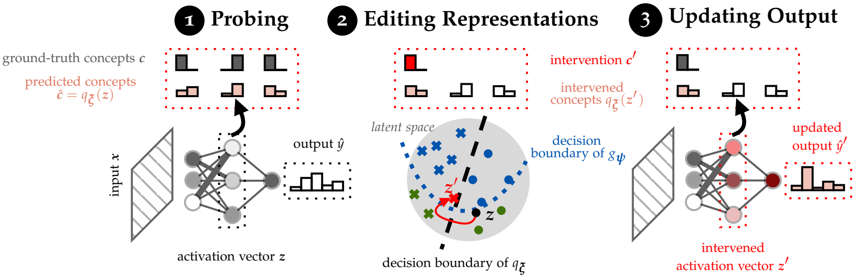

① Probing
② Editing Representations
③ Updating Output
ground-truth concepts c
predicted concepts $\hat{c} = q_\xi(z)$
input x
output $\hat{y}$
activation vector z
intervention c'
intervened concepts $q_\xi(z')$
latent space
decision boundary of g_ψ
z'
z
decision boundary of q_ξ
updated output $\hat{y}'$
intervened activation vector z'

FIGURE 6.1: Schematic summary of concept-based instance-specific interventions on a black-box neural network. Given concept values c' for an input x, our intervention procedure edits the network's activation vector z at an intermediate layer via the probing function q_ξ, returning an intervened activation vector z' coherent with the given concepts. The intervention results in an updated prediction $\hat{y}'$.

① The first and preliminary step is to train a multivariate **probing** function [293], [294], or probe for short, to map the network's intermediate representations to concept variables. Namely, using the given validation set, we train a probe q_ξ to predict concepts c_i from the representations $z_i = h_\phi(x_i)$ for $1 \leq i \leq N$: $\min_\xi \sum_{i=1}^{N} \sum_{k=1}^{K} \ell^{c_k}\left(q_\xi(z_i)_k, c_{i,k}\right)$, where, as in Section 5.2.2, $K \geq 1$ is the number of concept variables and ℓ^{c_k} denotes the loss function for the k-th concept. Note that $q_\xi(z_i)_k$ above corresponds to the k-th output of the probing function for the input z_i. An essential design choice we explore in our experiments (Section 6.3.2) is the (non)linearity of the function q_ξ. Upon this step, q_ξ is utilised to probe the representations and interpret them via predicted concepts $\hat{c} = q_\xi(x)$.

② At intervention time, we are given a data point (x, y) and concept values c'. Note that this c' could correspond to the ground-truth concept values c or reflect the beliefs of a domain expert interacting with the model. In the second step of the intervention procedure, we **edit the representation**, seeking an activation vector z' (i) similar to z and (ii) consistent with c' according to the previously learnt probing function q_ξ. Formally, z' is given by

$$\arg\min_{z'} d\left(z, z'\right), \text{ subject to } q_\xi\left(z'\right) = c', \tag{6.1}$$

where d is an appropriate distance function applied to the activation vectors. Throughout this chapter's experiments (Section 6.3), we utilise the Euclidean metric frequently applied to neural network representations, e. g.

see [358] and [359]. However, other functions may be leveraged, such as cosine dissimilarity.

Instead of the constrained problem in Equation 6.1, we resort to minimising a relaxed objective:

$$\min_{z'} d\left(z, z'\right) + \frac{\lambda}{K} \sum_{k=1}^{K} \ell^{c_k}\left(q_{\xi}\left(z'\right)_k, c'_k\right), \tag{6.2}$$

where the hyperparameter $\lambda > 0$ controls the tradeoff between the intervention's *validity*, i. e. the consistency of z' with the given concept values c' according to the probe, and *proximity* to the original activation vector z. We explore the influence of this hyperparameter empirically in the experiments (Section 6.3). Note the resemblance between Equation 6.2 and the definition of counterfactual explanations [12] (Definition 2.4.6). In our case, the optimisation problem is defined on the network's activations instead of the raw feature space and the constraint is provided by the concepts rather than the target. In practice, z' can be optimised using gradient descent for a single intervention or a batch thereof.

❸ Finally, the edited z' (Equation 6.2) can be fed into g_{ψ} to compute the **updated output** $\hat{y}' = g_{\psi}\left(z'\right)$, which could then be returned and displayed to the user. For example, if c' are the ground-truth concept values, we would ideally expect a decrease in the prediction error for the given data point (x, y).

Naturally, interventions performed on black-box models using our method are meaningful in so far as the activations of the neural network are correlated with the given high-level attributes and the probing function q_{ξ} can be trained to predict these attribute values accurately. Otherwise, edited representations (Equation 6.2) and updated predictions are likely to be spurious and may harm the model's performance.

6.2.1 *Assessing Intervention Effectiveness*

Conventionally, CBMs [22] and their extensions are evaluated empirically by plotting changes in test-set performance or error when intervening on concept subsets of varying sizes, e. g. as in Figure 5.6. Ideally, the model's test-set performance should improve when provided with more ground-truth attribute values. We will refer to this notion of intervention effectiveness as *intervenability*. Below, we define intervenability formally for the concept bottleneck and black-box models.

To begin with, let us consider a CBM $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\psi}}(h_{\boldsymbol{\phi}}(\boldsymbol{x})) = g_{\boldsymbol{\psi}}(\hat{\boldsymbol{c}})$. For CBMs, we define the intervenability as follows:

$$\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)}\left[\mathbb{E}_{\pi(\boldsymbol{c}')}\left[\ell^{y}\Big(\underbrace{f_{\boldsymbol{\theta}}(\boldsymbol{x})}_{\hat{y}=g_{\boldsymbol{\psi}}(\hat{\boldsymbol{c}})},y\Big) - \ell^{y}\Big(\underbrace{g_{\boldsymbol{\psi}}(\boldsymbol{c}')}_{\hat{y}'},y\Big)\right]\right], \tag{6.3}$$

where $p(\boldsymbol{x},\boldsymbol{c},y)$ is the joint distribution over the covariates, concepts, and targets, ℓ^{y} is the target prediction loss (Section 5.2.2), and $\pi(\boldsymbol{c}')$ denotes a distribution over edited concept values $\boldsymbol{c}'$. Thus, the effectiveness of interventions is quantified by the gap between the regular prediction loss and the loss attained after the intervention: the larger the gap between these values, the stronger the effect interventions have.

Note that Equation 6.3 can accommodate sophisticated intervention strategies, such as those studied by Sheth *et al.* [355] and Shin *et al.* [348]. An intervention strategy is specified via the distribution π, which may be conditioned on $\boldsymbol{x}$, $\hat{\boldsymbol{c}}$, $\boldsymbol{c}$, $\hat{y}$, or even y: $\pi(\boldsymbol{c}'|\boldsymbol{x},\hat{\boldsymbol{c}},\boldsymbol{c},\hat{y},y)$. The set of conditioning variables may vary across application scenarios. For brevity, we will use $\pi(\boldsymbol{c}')$ as a shorthand notation. Lastly, notice that, in practice, when performing human- or application-grounded evaluation [15], sampling from π should be replaced with the interventions performed by an expert. Throughout the current chapter, we will primarily concentrate on the simple random-subset and uncertainty-based strategies, which condition on $\boldsymbol{c}$ and $\hat{\boldsymbol{c}}$. Both intervention approaches are described in detail in Appendix D.1 (Algorithms D.1.1 and D.1.2).

Based on Equation 6.3 and the intervention procedure introduced above, we can now define the intervenability measure for black-box models.

Definition 6.2.1 [Intervenability]**.** For a black-box neural network $f_{\boldsymbol{\theta}}$ and intermediate layer given by $\langle g_{\boldsymbol{\psi}}, h_{\boldsymbol{\phi}}\rangle$, the intervenability is defined as

$$\begin{aligned}&\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)}\left[\mathbb{E}_{\pi(\boldsymbol{c}')}\left[\ell^{y}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big) - \ell^{y}\big(g_{\boldsymbol{\psi}}(\boldsymbol{z}'),y\big)\right]\right],\\ &\text{where } \boldsymbol{z}' \in \arg\min_{\tilde{\boldsymbol{z}}} d(\boldsymbol{z},\tilde{\boldsymbol{z}}) + \frac{\lambda}{K}\sum_{k=1}^{K}\ell^{c_k}\left(q_{\boldsymbol{\xi}}(\tilde{\boldsymbol{z}})_k, c_k'\right),\end{aligned} \tag{6.4}$$

where $q_{\boldsymbol{\xi}}$ is a probe trained to predict $\boldsymbol{c}$ based on the activations $\boldsymbol{z} = h_{\boldsymbol{\phi}}(\boldsymbol{x})$.

Note that in the first line of Equation 6.4, edited representations $\boldsymbol{z}'$ depend on $\boldsymbol{c}'$, as defined by the second line, corresponding to Equation 6.2. In practice, to obtain an empirical estimate of the quantity in Equation 6.4 and other measures and loss functions introduced hereon, the outer expectation w.r.t. $p\left(\boldsymbol{x}, \boldsymbol{c}, y\right)$ is approximated by averaging over the given dataset, while the expectation w.r.t. $\pi\left(\boldsymbol{c}'\right)$ is approximated using the MC method.

Thus, the intervenability measure (Definition 6.2.1) combined with the probing function can be used to evaluate the interpretability of a black-box predictive model and help understand whether (i) learnt representations capture information about given high-level attributes and whether (ii) the network utilises these attributes and can be interacted with. However, a black-box model does not always need to be intervenable. For instance, when the given concept set is not predictive of the target variable, the black box trained using supervised learning should not and probably would not rely on the concepts. On the other hand, if the model's representations are nearly perfectly correlated with the attributes, providing the ground truth should not significantly impact the target prediction loss.

6.2.2 *Fine-tuning for Intervenability*

Observe that the intervenability measure is differentiable and, therefore, can be explicitly maximised using (minibatch) gradient descent. We hypothesise that fine-tuning for intervenability will reinforce the model's reliance on the high-level attributes and have a regularising effect. In the current subsection, we introduce this fine-tuning procedure with the detailed pseudocode provided in Algorithm 3.

To fine-tune an already trained black-box model $f_{\boldsymbol{\theta}}$, we combine the target prediction loss with the weighted intervenability term and consider the following optimisation problem:

$$\begin{aligned} \min_{\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{z}'} \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{c}, y)} & \left[\mathbb{E}_{\pi(\boldsymbol{c}')} \left[(1-\beta)\, \ell^{y}\left(g_{\boldsymbol{\psi}}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right), y\right) + \beta \ell^{y}\left(g_{\boldsymbol{\psi}}\left(\boldsymbol{z}'\right), y\right) \right] \right], \\ \text{subject to } \; & \boldsymbol{z}' \in \arg\min_{\tilde{\boldsymbol{z}}} d\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}\right) + \frac{\lambda}{K} \sum_{k=1}^{K} \ell^{c_k}\left(q_{\boldsymbol{\xi}}\left(\tilde{\boldsymbol{z}}\right)_k, c'_k\right), \end{aligned} \tag{6.5}$$

where $\beta \in (0, 1]$ is the weight of the intervenability term. For simplicity, we treat the probe's parameters $\boldsymbol{\xi}$ as fixed. However, since the outer optimisation problem in Equation 6.5 is defined w.r.t. parameters $\boldsymbol{\phi}$, ideally, the probe needs to be optimised at the third, inner-most level of the problem.

Algorithm 3: Fine-tuning for Intervenability

```
Input: A trained black box $f_{\boldsymbol{\theta}} = \langle g_{\boldsymbol{\psi}}, h_{\boldsymbol{\phi}} \rangle$; probe $q_{\boldsymbol{\xi}}$; concept prediction loss functions
       $\ell^{c_k}$ for $1 \leq k \leq K$; target prediction loss $\ell^{y}$; validation set
       $\mathscr{D}_{\text{valid}} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)\}_{i=1}^{N}$; intervention strategy $\pi$; distance function $d$;
       hyperparameter value $\lambda > 0$; maximum number of steps $E_I \geq 1$ for the
       intervention procedure; convergence criterion parameter $\varepsilon_I > 0$; learning rate
       $\eta_I > 0$; number of fine-tuning epochs $E \geq 1$; minibatch size $M \geq 1$; learning
       rate $\eta > 0$
Output: Fine-tuned model

Train the probing function $q_{\boldsymbol{\xi}}$ on the validation set, i.e. let
  $\boldsymbol{\xi} \leftarrow \arg\min_{\boldsymbol{\xi}'} \sum_{i=1}^{N} \sum_{k=1}^{K} \ell^{c_k}\left(q_{\boldsymbol{\xi}'}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}_i\right)\right)_k, c_{i,k}\right)$

for $e = 0$ to $E - 1$ do
    Randomly split $\{1, \ldots, N\}$ into minibatches of size $M$
    for minibatch $\mathscr{B} \subseteq \{1, \ldots, N\}$ do
        for $i \in \mathscr{B}$ do
            Let $\boldsymbol{z}_i \leftarrow h_{\boldsymbol{\phi}}\left(\boldsymbol{x}_i\right)$
            Let $\hat{y}_i \leftarrow g_{\boldsymbol{\psi}}\left(\boldsymbol{z}_i\right)$
            Let $\hat{\boldsymbol{c}}_i \leftarrow q_{\boldsymbol{\xi}}\left(\boldsymbol{z}_i\right)$

            Sample $\boldsymbol{c}'_i \sim \pi$
            Initialise $\boldsymbol{z}'_i = \boldsymbol{z}_i$, $\boldsymbol{z}'_{i,\text{old}} = \boldsymbol{z}_i + \varepsilon_I$, and $e_I = 0$
        end

        while $\sum_{i \in \mathscr{B}} \left\| \boldsymbol{z}'_i - \boldsymbol{z}'_{i,\text{old}} \right\|_1 \geq \varepsilon_I$ and $e_I < E_I$ do
            for $i \in \mathscr{B}$ do
                Update $\boldsymbol{z}'_{i,\text{old}} \leftarrow \boldsymbol{z}'_i$
                Update
                    $$\boldsymbol{z}'_i \leftarrow \boldsymbol{z}'_i - \eta_I \nabla_{\boldsymbol{z}'_i}\left[d(\boldsymbol{z}_i, \boldsymbol{z}'_i) + \frac{\lambda}{K}\sum_{k=1}^{K} \ell^{c_k}\left(q_{\boldsymbol{\xi}}\left(\boldsymbol{z}'_i\right)_k, c'_{i,k}\right)\right]$$
            end
            Update $e_I \leftarrow e_I + 1$
        end

        Let $\hat{y}'_i \leftarrow g_{\boldsymbol{\psi}}\left(\boldsymbol{z}'_i\right)$ for $i \in \mathscr{B}$

        Update $\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} - \eta \nabla_{\boldsymbol{\psi}} \sum_{i \in \mathscr{B}} \ell^{y}\left(\hat{y}'_i, y_i\right)$
    end
end

return $f_{\boldsymbol{\theta}}$
```

To avoid computationally costly trilevel optimisation, we consider a special case of Equation 6.5 under $\beta = 1$. The problem simplifies to

$$\begin{aligned} &\min_{\boldsymbol{\psi}, \boldsymbol{z}'} \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)} \left[\mathbb{E}_{\pi(\boldsymbol{c}')} \left[\ell^{y} \left(g_{\boldsymbol{\psi}} \left(\boldsymbol{z}' \right), y \right) \right] \right], \\ &\text{subject to } \; \boldsymbol{z}' \in \arg\min_{\tilde{\boldsymbol{z}}} d\left(\boldsymbol{z}, \tilde{\boldsymbol{z}} \right) + \frac{\lambda}{K} \sum_{k=1}^{K} \ell^{c_k} \left(q_{\boldsymbol{\xi}} \left(\tilde{\boldsymbol{z}} \right)_k, c'_k \right). \end{aligned} \tag{6.6}$$

In Equation 6.6, the parameters of $h_{\boldsymbol{\phi}}$ do not need to be optimised, and, hence, the probing function can be left fixed, as activations $\boldsymbol{z}$ are not affected by the fine-tuning. We consider this case to (i) computationally simplify the problem and (ii) keep the network's representations unchanged after fine-tuning for generalisation across other downstream tasks.

In practice, fine-tuning is performed by intervening on minibatches of data points, as shown in Algorithm 3 (lines 12–19). Note that this implementation corresponds to the special case of $\beta = 1$ with the simplified loss function. Hence, the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are not updated during gradient descent (lines 2–22).

6.3 EXPERIMENTAL SETUP

Having introduced novel methods for concept-based intervention on black-box neural networks and explicit fine-tuning to improve intervention effectiveness, we now describe the experimental setup of this chapter. In particular, we explain the benchmarking datasets, baseline methods, and evaluation metrics utilised for empirical comparison.

6.3.1 *Datasets*

We evaluate the proposed methods on synthetic and real-world classification benchmarks summarised in Table 6.1. All datasets are divided according to the 60%-20%-20% train-validation-test split. Fine-tuning is performed solely on the validation data, and evaluation is conducted on the test set.

For controlled experiments, we adapt the nonlinear tabular **synthetic** dataset from Chapter 5 (Section 5.3.1 and Appendix C.1) to the single-view classification scenario. Additionally, similar to Shin *et al.* [348], we consider a few distinct data-generating mechanisms summarised graphically in Figure 6.2. We refer to these as *bottleneck*, *confounder*, and *incomplete*. In

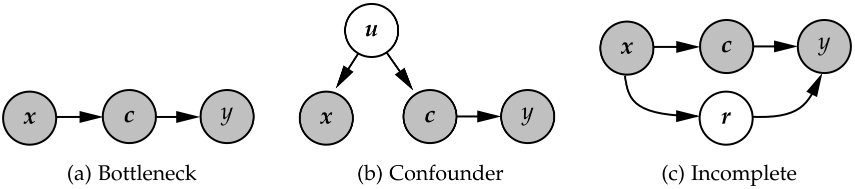
Dataset	Data type	N	p	K
Synthetic	Tabular	50000	1500	30
AwA	Image	37322	224×224	85
CheXpert	Image	49408	224×224	13
MIMIC-CXR	Image	54276	224×224	13

TABLE 6.1: A summary of datasets. After any filtering or preprocessing, N denotes the total number of data points, p is the input dimensionality, and K is the number of concept variables.

brief, the *bottleneck* scenario (Figure 6.2a) directly matches the inference graph of the vanilla CBM [22]. For the *confounder* (Figure 6.2b), $\boldsymbol{c}$ and $\boldsymbol{x}$ are generated by an unobserved confounder, and y is generated by $\boldsymbol{c}$. Lastly, the *incomplete* setting (Figure 6.2c) is similar to the insufficient concept set scenario from Section 5.2.3. In this instance, $\boldsymbol{c}$ does not fully explain y and unexplained variance is modelled as a latent variable along the residual path from $\boldsymbol{x}$ to y. Appendix D.2 provides a detailed description of data-generating procedures for the three scenarios. In this chapter, unless explicitly mentioned, we primarily concentrate on the simplest *bottleneck* setting.

As in the previous chapter, we also consider the **Animals with Attributes 2** natural image dataset [112], [331] comprising animal images accompanied by binary attributes and species labels. In contrast to Section 5.3.1, we utilise the original dataset without cropping and constructing multiple views.

A higher-complexity biomedical problem we explore in this chapter is chest radiograph classification. In particular, we test the methods on the publicly available **CheXpert** dataset [360] from Stanford Hospital. It includes



(a) Bottleneck (b) Confounder (c) Incomplete

FIGURE 6.2: Data-generating mechanisms for the synthetic dataset summarised as graphical models. Each node corresponds to a random variable. Shaded nodes denote observed variables.

over 220000 chest radiographs from 65240 patients. X-ray images are accompanied by 14 binary attributes extracted from radiologist reports using the CheXpert labeller [360].

Another similar benchmark is the **Medical Information Mart for Intensive Care Chest X-ray (MIMIC-CXR)** database [361]. MIMIC-CXR comprises more than 370000 images associated with 227835 radiographic studies conducted at the Beth Israel Deaconess Medical Center, Boston, MA, USA, involving 65379 patients. As in the CheXpert, the labeller was employed to extract the set of 14 binary labels from accompanying text reports.

For both chest X-ray datasets, we closely follow the setup proposed by Chauhan *et al.* [356]. We designate the "finding/no findings" attribute as the target variable for classification and utilise the remaining labels as concept variables. For the purity of the experiment, we remove all the samples with uncertain labels and discard multiple visits of the same patient, keeping only the last acquired recording per subject. As part of the standard preprocessing routine, all images are cropped to the square aspect ratio and rescaled to 224×224 pixels. In addition, image augmentations are applied during training to avoid overfitting.

6.3.2 *Baselines and Ablations*

Our experiments compare several neural network model variants and fine-tuning methods to improve intervention effectiveness. Below, we summarily introduce all techniques and ablation studies. In Appendix D.3, we provide additional details on network architectures.

Firstly, we train a standard **black-box** neural network without concept knowledge, i. e. on the dataset of tuples $\{(\boldsymbol{x}_i, y_i)\}_i$. We utilise the concept-based intervention technique introduced in Section 6.2 by (i) training a probing function on the validation set to predict concepts and (ii) editing the network's activations (Equation 6.2). In our ablation studies, we additionally explore the influence of the hyperparameter λ and the nonlinearity of the probe on the intervention effectiveness. Furthermore, we compare two intervention strategies: random-subset and uncertainty-based (Appendix D.1).

As an interpretable baseline, we consider the vanilla **CBM** [22]. For brevity, we only report results for the joint optimisation procedure (Equation 5.6), as we did not observe significant differences across training techniques.

In addition to black boxes and CBMs, we investigate the impact of fine-tuning. We define several common-sense baseline techniques. The first approach is fine-tuning via a probing function trained to predict concept variables (denoted in short by **Fine-tuned, MT**). This technique is similar to *multitask* learning with hard weight sharing [362]. In particular, the optimisation problem is

$$\min_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\xi}} \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)} \left[\ell^{y}\left(g_{\boldsymbol{\psi}}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right), y\right) + \frac{\alpha}{K} \sum_{k=1}^{K} \ell^{c_k}\left(q_{\boldsymbol{\xi}}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right)_k, c_k\right) \right], \tag{6.7}$$

where $\alpha > 0$ controls the tradeoff between target and concept prediction loss terms.

Another approach is fine-tuning by *appending* concepts to the network's activations (**Fine-tuned, A**). For binary variables, at test time, we set unknown concept values to 0.5. Furthermore, to prevent overfitting and handle missingness, randomly chosen concept variables are masked during training. Formally, this fine-tuning technique can be summarised as

$$\min_{\tilde{\boldsymbol{\psi}}} \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)} \left[\ell^{y}\left(\tilde{g}_{\tilde{\boldsymbol{\psi}}}\left(\left[h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right), \boldsymbol{c}\right]\right), y\right) \right] \tag{6.8}$$

where $\tilde{g}_{\tilde{\boldsymbol{\psi}}}$ takes as input concatenated activation and concept vectors. Notably, in contrast to Equation 6.7, in this technique, parameters $\boldsymbol{\phi}$ remain fixed during fine-tuning.

As a strong baseline, we train a CBM *post hoc* on the black-box network's representations (**Post hoc CBM**). This method resembles approaches by Yüksekgönül, Wang, and Zou [343] and Oikarinen *et al.* [351] (Section 6.1.1) but features a few technical deviations taken to facilitate a fairer and more direct comparison with our method. In particular, we perform *sequantial* optimisation by first predicting the concepts from the backbone's representations and then training the target model:

$$\begin{aligned} \hat{\boldsymbol{\xi}} &= \arg\min_{\boldsymbol{\xi}} \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)} \left[\sum_{k=1}^{K} \ell^{c_k}\left(q_{\boldsymbol{\xi}}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right)_k, c_k\right) \right], \\ \hat{\boldsymbol{\psi}} &= \arg\min_{\boldsymbol{\psi}} \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c},y)} \left[\ell^{y}(g_{\boldsymbol{\psi}}(q_{\hat{\boldsymbol{\xi}}}(h_{\boldsymbol{\phi}}(\boldsymbol{x}))), y) \right]. \end{aligned} \tag{6.9}$$

We omit the inclusion of a residual channel in the *post hoc* CBM, as proposed by Yüksekgönül, Wang, and Zou [343], since we observe little improvement from this step in practice.

Lastly, the technique of primary interest in these experiments is fine-tuning explicitly for the intervenability (Equations 6.5–6.6, Algorithm 3; **Fine-tuned, I**). During fine-tuning, we utilise the random-subset intervention strategy (Algorithm D.1.1), performing interventions based on the 50% of concept variables sampled uniformly at random.

6.3.3 *Evaluation Metrics*

To compare the models and fine-tuning methods mentioned in Section 6.3.2, we conduct interventions and analyse target prediction performance under varying sizes of the sets of concepts intervened on. To this end, we report changes in AUROCs and AUPRs similar to those shown in Figure 5.6 of Chapter 5. As a sanity check, we also examine target and concept prediction AUROCs, AUPRs, and Brier scores *without* interventions. Concept prediction performance is evaluated either directly on the bottleneck layer (for *ante* and *post hoc* CBMs) or by fitting a multivariate logistic regression model on the activations of the appropriate layer.

6.4 RESULTS

We now turn to our experimental findings. We begin with the proof-of-concept experiments on the synthetic dataset. Subsequently, we benchmark models and fine-tuning techniques on natural images, providing ablation studies on the impact of essential hyperparameters and design choices. At the end of this section, we investigate the application of our methods to deep chest X-ray classification models. In addition to assessing concept-based intervention effectiveness, as explained in Section 6.3.3, we evaluate models' test-set performance without interventions. These results are reported in Table 6.2 across all datasets and will be discussed in the corresponding subsections.

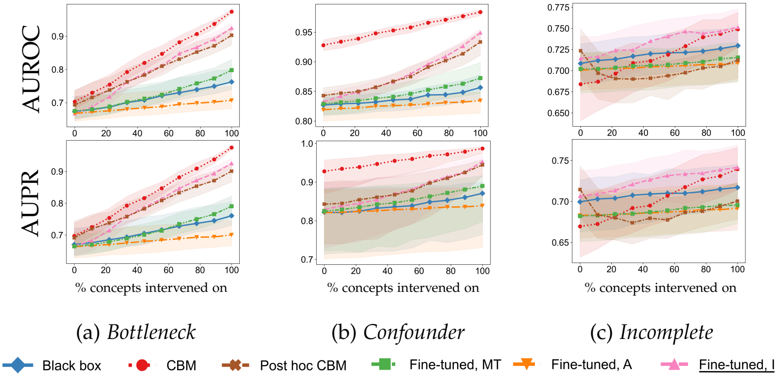
6.4.1 *Results on Synthetic Data*

As mentioned in Section 6.3.1, we generate synthetic nonlinear tabular data using three different procedures representing plausible mechanisms (Figure 6.2, Appendix D.2).

Figure 6.3 shows intervention results obtained across ten independent simulations. Across all three scenarios, we observe that, in principle, the proposed intervention procedure can improve the predictive performance

Dataset	**Model**	**Concepts**			**Target**		
		AUROC	**AUPR**	**Brier**	**AUROC**	**AUPR**	**Brier**
Synthetic	Black box	0.716±0.018	0.710±0.017	0.208±0.006	0.686±0.043	0.675±0.046	0.460±0.003
	CBM	**0.837±0.008**	**0.835±0.008**	*0.196±0.006*	**0.713±0.040**	**0.700±0.038**	*0.410±0.012*
	Post hoc CBM	0.714±0.017	0.707±0.018	0.207±0.009	*0.707±0.049*	*0.698±0.048*	**0.285±0.015**
	Fine-tuned, A	—	—	—	0.682±0.047	0.668±0.046	0.470±0.004
	Fine-tuned, MT	*0.784±0.013*	*0.780±0.014*	**0.186±0.006**	0.687±0.046	0.668±0.043	0.471±0.003
	Fine-tuned, I	0.716±0.018	0.710±0.017	0.208±0.006	0.695±0.051	0.685±0.051	**0.285±0.014**
AwA	Black box	0.991±0.002	*0.979±0.006*	0.027±0.006	*0.996±0.001*	0.926±0.020	0.199±0.038
	CBM	*0.993±0.001*	*0.979±0.002*	*0.025±0.001*	0.988±0.001	0.892±0.005	0.234±0.009
	Post hoc CBM	0.992±0.002	0.976±0.005	*0.025±0.005*	*0.996±0.001*	*0.929±0.018*	**0.170±0.033**
	Fine-tuned, A	—	—	—	*0.996±0.001*	**0.938±0.016**	**0.170±0.036**
	Fine-tuned, MT	**0.994±0.002**	**0.985±0.004**	**0.022±0.005**	**0.997±0.001**	**0.938±0.017**	*0.178±0.038*
	Fine-tuned, I	0.991±0.002	*0.979±0.005*	0.027±0.006	*0.996±0.001*	0.925±0.020	0.195±0.040
CheXpert	Black box	0.665±0.003	0.257±0.003	*0.097±0.001*	0.785±0.011	0.911±0.006	0.305±0.009
	CBM	**0.723±0.005**	**0.322±0.003**	0.116±0.001	*0.786±0.009*	0.919±0.006	0.375±0.013
	Post hoc CBM	0.597±0.007	0.222±0.003	0.103±0.001	**0.819±0.008**	**0.939±0.004**	*0.207±0.005*
	Fine-tuned, A	—	—	—	0.749±0.008	0.891±0.005	0.329±0.013
	Fine-tuned, MT	*0.684±0.003*	*0.275±0.003*	**0.094±0.001**	0.768±0.019	0.901±0.012	0.297±0.012
	Fine-tuned, I	0.668±0.004	0.257±0.003	*0.097±0.001*	**0.819±0.009**	*0.938±0.004*	**0.201±0.007**
MIMIC-CXR	Black box	0.743±0.006	0.170±0.004	*0.046±0.001*	0.789±0.006	0.706±0.009	0.444±0.003
	CBM	*0.744±0.006*	**0.224±0.003**	0.053±0.001	0.765±0.007	0.699±0.006	0.427±0.003
	Post hoc CBM	0.707±0.006	0.154±0.006	*0.046±0.001*	*0.801±0.006*	*0.727±0.008*	**0.301±0.005**
	Fine-tuned, A	—	—	—	0.773±0.009	0.665±0.013	0.459±0.004
	Fine-tuned, MT	**0.748±0.008**	*0.187±0.003*	**0.045±0.001**	0.785±0.006	0.696±0.009	0.450±0.008
	Fine-tuned, I	*0.744±0.005*	0.172±0.005	*0.046±0.001*	**0.808±0.007**	**0.733±0.009**	*0.314±0.015*

Table 6.2: Concept and target prediction performance *without interventions*. For black-box models, concepts are predicted via a linear probe, and their prediction metrics are averaged. The synthetic dataset was generated under the *bottleneck* scenario. Best results are reported in **bold**, second best are in *italics*.



(a) *Bottleneck* (b) *Confounder* (c) *Incomplete*

Figure 6.3: Intervention effectiveness w.r.t. target prediction AUROC (*top*) and AUPR (*bottom*) on the synthetic data generated under the (a) *bottleneck*, (b) *confounder*, and (c) *incomplete* scenarios. Bold lines correspond to medians, and confidence bands are given by interquartile ranges.

of a black-box neural network. However, as expected, interventions are considerably more effective in CBMs than in untuned black-box classifiers: the former exhibit a steeper increase in AUROC and AUPR, given more ground-truth concept values.

Generally, models explicitly fine-tuned for intervenability (Fine-tuned, I) significantly improve over the original classifier, achieving intervention curves comparable or better than those of the CBM for the *bottleneck* (Figure 6.3a) and *incomplete* (Figure 6.3c) settings. Importantly, under an incomplete concept set, black-box classifiers are expectedly superior to the *ante hoc* CBM, and fine-tuning for intervenability improves intervention effectiveness while maintaining the performance gap.

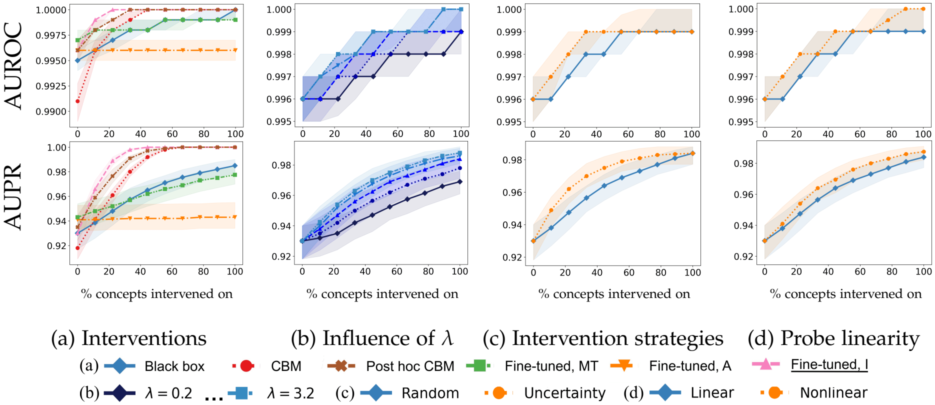
Other fine-tuning strategies (Fine-tuned, MT; Fine-tuned, A) are either less effective or harmful, leading to a lower increase in AUROC and AUPR. Lastly, CBMs trained *post hoc* perform well in the *bottleneck* and *confounder* scenarios, being only slightly less intervenable than the model fine-tuned using our approach. However, in the *incomplete* setting, interventions on the *post hoc* CBM hurt the model's performance, leading to a noticeable decrease in target prediction AUROC and AUPR.

Let us now inspect test-set performance *without* interventions reported in Table 6.2. For the concept prediction, expectedly, CBM outperforms black-box models, except for those fine-tuned with the MT loss. However, all models attain somewhat comparable AUROCs and AUPRs at the target prediction. Interestingly, fine-tuning for intervenability results in better-calibrated probabilistic predictions with visibly lower Brier scores. Thus, while the proposed fine-tuning technique does not considerably affect the vanilla concept or target prediction, as intended, it improves black-box intervention effectiveness, facilitating human–model interaction.

6.4.2 *Results on Natural Images*

To corroborate the findings reported above, we explore the natural image AwA dataset. In addition to the method comparison, we utilise this simple benchmark for ablation experiments to investigate the role of important hyperparameters and design choices embedded in our *post hoc* intervention procedure (Section 6.2).

We compare intervention effectiveness among black-box models and *ante* and *post hoc* CBMs in Figure 6.4a. Similar to the synthetic dataset, AwA is a simple classification benchmark with class-wide concepts that help predict the target variable. Therefore, CBMs trained *ante* and *post hoc* are



(a) Interventions (b) Influence of λ (c) Intervention strategies (d) Probe linearity

FIGURE 6.4: Intervention results on the AwA dataset w.r.t. target AUROC (*top*) and AUPR (*bottom*). (a) Comparison among the models and fine-tuning techniques considered. (b) Intervention results for the untuned black-box model under varying values of $\lambda \in \{0.2, 0.4, 0.8, 1.6, 3.2\}$. **Darker** and **lighter** colours correspond to lower and higher λ-values, respectively. (c) Comparison between **random-subset** and **uncertainty-based** intervention strategies. (d) Comparison between interventions under **linear** and **nonlinear** probing functions.

highly performant and intervenable. In agreement with the previous findings, our fine-tuning approach enhances the performance of black-box models, resulting in steeper curves than those of CBMs. Overall, the simplicity of the dataset leads to the relatively saturated AUROCs and AUPRs across all methods. This observation is also confirmed by Table 6.2. In particular, *without* interventions, all models are successful, likely due to the large dataset size and relative simplicity of the classification task.

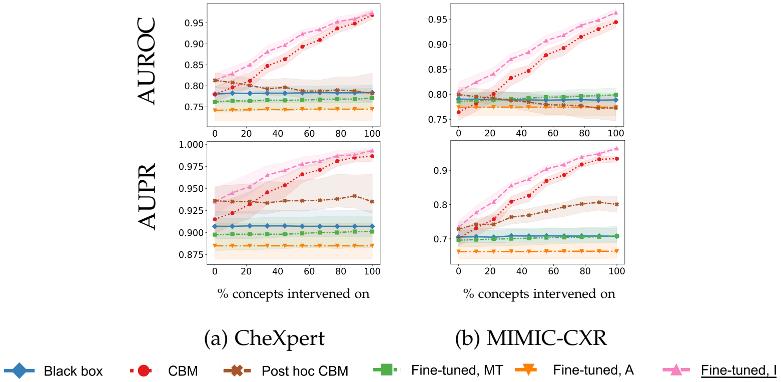
As stated above, we also perform ablation studies on the intervention procedure applied to *untuned* black boxes. The results are visualised in Figures 6.4b–6.4d. First, we vary the λ-parameter from Equation 6.4, weighting the concept-loss term. Figure 6.4b shows that interventions are effective across all values of λ. Expectedly, higher hyperparameter values yield more effective interventions.

In Figure 6.4c, we compare two intervention strategies: randomly selecting a concept subset (Random) and prioritising uncertain concepts (Uncertainty) to intervene on [348] (Appendix D.1). The intervention strategy has a clear impact on the performance increase, with the uncertainty-based approach yielding a steeper improvement. Finally, Figure 6.4d compares linear and nonlinear probes. Similar to the effect of the uncertainty-based strategy, intervening via a nonlinear function leads to a significantly higher target prediction performance increase.

6.4.3 *Application to Deep Chest X-ray Classifiers*

Lastly, we comment on our experiments on two chest radiograph datasets with classifiers predicting pathological findings. Figure 6.5 presents the comparison among the models and fine-tuning techniques on the CheXpert (Figure 6.5a) and MIMIC-CXR (Figure 6.5b) data.

By contrast to Sections 6.4.1 and 6.4.2, in both chest X-ray datasets, *untuned* black-box neural networks are not intervenable. However, after fine-tuning for intervenability, the model's predictive performance and effectiveness of interventions improve visibly and even surpass those of the *ante hoc* CBM. Given the challenging nature of these real-world datasets with instance-level concept labels, representations learnt by black-box neural networks may not be strongly reliant on the attributes. Furthermore, on average, CBMs do not outperform black-box models, unlike in simpler synthetic benchmarks (Figures 6.3a–6.3b). Lastly, *post hoc* CBMs exhibit a behaviour similar to that on the synthetic dataset with *incomplete*



(a) CheXpert (b) MIMIC-CXR

FIGURE 6.5: Intervention results w.r.t. target AUROC (*top*) and AUPR (*bottom*) on the (a) CheXpert and (b) MIMIC-CXR datasets.

concepts (Figure 6.3c): interventions have no or even an adverse effect on the predictive performance.

Without interventions, CBMs exhibit better performance at concept prediction, while fine-tuned models and *post hoc* CBMs outperform them at the target classification (Table 6.2). Similar to the experiments on the synthetic data, alongside AUROCs and AUPRs, fine-tuning for intervenability leads to improved Brier scores.

6.5 DISCUSSION

Having described our empirical findings, let us turn to the general discussion of the contributions and relevance of the current chapter. Below, we briefly summarise the proposed methods and their relation to the broader context of the thesis, compare our techniques to the closely related literature, and comment on the empirical contributions and experimental insights. We conclude by reflecting on the limitations of the methods and experimental setup and discuss potential directions for future work.

Similar to Chapter 5, we have treated the problem of explaining predictive models via high-level concept variables (Section 2.4.3). Instead of explicitly incorporating concepts into the neural network's architecture, as done by concept bottleneck models [22], we investigate the interaction with a neural network using high-level attributes *post hoc*. In particular, we propose a procedure for instance-specific concept-based interventions on the network's representations (Figure 6.1). Such interventions allow users to affect the network's final output by providing a set of attribute values. Our approach is faithful to the network's (i) architecture and (ii) representations, as it leverages an *external* probing function [293], [294] to interpret and edit activation vectors. Thus, our technique does not rely on the restrictive bottleneck layer. Beyond concept-based models and explanation techniques, the proposed intervention procedure builds on counterfactual explanations (Section 2.4.5) as defined by Wachter, Mittelstadt, and Russell [12]. Specifically, the optimisation problem for representation editing we consider (Equation 2.13) is analogous to the one for counterfactual explanations (Definition 2.4.6).

In addition, we have formalised the model's *intervenability* as a measure of the effectiveness of concept-based interventions (Section 6.2.1) for both CBMs (Equation 6.3) and black-box neural networks (Equation 6.4). In brief, intervenability corresponds to the expected gap in target prediction loss without and under interventions. While, in practice, not all models need to be intervenable, this measure allows evaluating and ranking black-box neural networks, which might be otherwise comparable w.r.t. predictive performance. Beyond the concept-based prediction setting, our intervenability measure may have utility in the evaluation of deep latent variable models [105], [232], [233], such as the ones introduced in Chapter 4. With a few adaptations, intervenability can help gauge the relationship between the concepts and the model's latent space and reconstructions.

Another technical contribution of this chapter is the procedure for the *explicit* fine-tuning of black-box models to improve their intervenability (Section 6.2.2). Given a labelled and annotated validation set, our method maximises the intervenability measure combined with the target prediction loss (Equation 6.5, Algorithm 3). Thus, fine-tuning for intervenability is meant to reinforce the model's reliance on the concept variables, potentially, without altering backbone representations (cf. Equations 6.5 and 6.6).

To summarise from a broader perspective, this chapter has explored tradeoffs between interpretability, intervenability, and performance in black-box predictive models. The proposed procedures facilitate effective human–model interaction while requiring a moderately sized *validation* set with concept labels for probing and fine-tuning.

Several lines of literature are closely related to the technical problems and methods introduced in this chapter. Naturally, the intuition behind concept-based interventions and intervenability is inspired by concept bottleneck models [22], [111], [112] and their ability to be interacted with (Equation 5.2).

In contrast to CBMs, our intervention and fine-tuning techniques are applicable to black-box neural networks *trained* without *concept* knowledge. However, a validation set is still needed for probing and fine-tuning, albeit considerably smaller than the training set. Furthermore, as discussed in Chapter 5, vanilla CBMs suffer from poorer performance under incomplete concept sets. Since our approach is faithful to the network's architecture and does not introduce a bottleneck layer, it makes no assumptions about the sufficiency of concepts.

In interpreting the network's intermediate representations or activation vectors, we capitalise on the previous research on probing [294] and concept activation vectors [138] (Section 2.4.3). In particular, the probing function is a crucial building block of the interventions, intervenability measure, and fine-tuning procedure. However, in contrast to the previous works, which focus on assessing correlations between representations and concept variables, this chapter leverages probes for *model editing* [363].

Similarly, we can draw a relation between concept-based interventions and conceptual counterfactual explanations [353]. As discussed in Section 6.1.1, CCEs identify a combination of concept variables that induce a change of the predicted label $\hat{y}$. By contrast, interventions are derived from a different optimisation problem that does not involve the model's final output (Equation 6.2). Thus, instead of seeking sparse sets of concept variables to explain the classifier's decision boundary, our aim is to perturb representations consistently with the *given* concept values.

Finally, the most closely related methods are *post hoc* and label-free CBMs [343], [351], which convert black-box models into CBMs *post hoc* by training concept and target prediction modules on frozen backbone representations. Furthermore, these works investigate using VLMs to label concepts based on verbal prompts and images. While the latter aspect is beyond the scope of this thesis, in principle, all techniques introduced in this chapter can be readily applied to VLM-labelled data. An essential advantage of our *post hoc* concept-based interventions and fine-tuning for intervenability over the approaches by Yüksekgönül, Wang, and Zou [343] and Oikarinen *et al.* [351] is faithfulness to the network's architecture. Since *post hoc* and label-free CBMs introduce an explicit bottleneck, both suffer from the inherent limitations of the *ante hoc* approach. Our empirical findings corroborate this claim, with *post hoc* CBMs having poor intervention effectiveness in more complex scenarios (Figures 6.3c and 6.5).

In addition to pure methodological contributions, we have provided a careful step-by-step empirical evaluation of our techniques (Section 6.4) on synthetic tabular and natural and medical image datasets (Section 6.3.1). Our experiments demonstrate that black-box neural network models trained without explicit concept knowledge are, in principle, intervenable, specifically, on synthetic and natural image benchmarks (Sections 6.4.1–6.4.2). Furthermore, explicitly fine-tuning for intervenability improves the effectiveness of concept-based interventions, bringing black-box models on par with *ante hoc* CBMs. Additionally, we have introduced a few common-sense fine-tuning baselines that perform visibly worse than the proposed technique, highlighting the need for explicitly including the intervenability measure in the loss function (Equation 6.5).

With regard to biomedical and healthcare applications, we have tested our techniques on chest X-ray classification (Section 6.4.3) using publicly available CheXpert [360] and MIMIC-CXR [361] datasets. In this setting, black-box classifiers are not directly intervenable (Figure 6.5). Nevertheless, fine-tuning alleviates this artefact. In contrast to the synthetic data, on medical images, fine-tuned models surpass CBMs in the test-set target prediction performance while featuring comparable intervention effectiveness.

In the broader research landscape, this chapter contributes to the notion of ML model understanding through editing [24] and facilitation of interaction between end users and predictive models. In the biomedical and healthcare domains, human–model interaction may become an instrumental feature in effectively combining human expertise and ML-based decision support systems, for instance, in medical image interpretation tasks, such as the ones studied in this and previous chapters. To the best of our knowledge, this work is one of the relatively few steps in that direction.

6.5.1 *Limitations*

Our methods and experimental setup have apparent limitations that warrant further discussion. A fundamental drawback of most concept-based models and explanation techniques is the need for annotated data at some step in the model's "life cycle". Similarly, our concept-based intervention and fine-tuning procedures utilise concept labels to train probing functions. As mentioned above, previous works [343], [351] have alleviated this limitation using VLMs. Additional empirical investigation is necessary to demonstrate that our techniques are likewise effective on the concepts discovered with the aid of multimodal models.

Another limitation of our methods is their computational complexity. Although concept-based interventions can be optimised in a batched manner, fine-tuning for intervenability requires bi- or even trilevel optimisation (Equation 6.5, Algorithm 3). In addition, the procedure is sensitive to the dimensionality of the layer intervened on, the number of given concept variables, and the design of the probing function.

Our current experiments serve primarily as a proof of concept, and most datasets explored are relatively simple (Section 6.3.1). For instance, the AwA dataset features only class-wide concept labels without variability across individual data points. In both chest radiograph datasets, concepts are *directly* related to the target variable and likely do not capture all clinically relevant information from images. Thus, a richer and more realistic healthcare or biomedical benchmark would be a valuable extension to the current setup.

Lastly, we restricted our analysis to jointly optimised CBMs (Equation 5.6) and random-subset and uncertainty-based intervention strategies (Appendix D.1). While we do not expect qualitatively different findings, a thorough investigation across training procedures and intervention strategies would aid the replicability of our results and help identify potential failure modes of our methods.

6.5.2 *Future Work*

In addition to addressing the abovementioned limitations, we describe avenues for future research and improvements. Since, for computational and practical reasons, we have simplified the optimisation problem behind our fine-tuning method (cf. Equations 6.5 and 6.6), it would be interesting to investigate the general formulation where all model and probe parameters are fine-tuned end-to-end. Furthermore, Algorithm 3 assumes being given a single fixed intervention strategy. We hypothesise that further improvement could come from *learning* an optimal strategy next to fine-tuning.

Another general direction for future work, also evident from Section 6.5.1, is the extension of the current experimental setup. In particular, building on Chapter 5, we could explore multiview and multimodal learning scenarios [32], [33] and investigate the utility of concept-based interventions and intervenability-based fine-tuning in black-box models trained on the pediatric appendicitis dataset (Section 5.3.2). Our evaluation has focused on simple MLP and CNN architectures (Appendix D.3). Therefore, investigating larger backbones is another direction for improvement. Lastly, beyond predictive modelling, we would like to explore the use of the intervenability measure (Equation 6.4) to evaluate deep latent variable models [105], [232], [233] (Section 4.1.2).

6.6 SUMMARY

This chapter studied the problem of instance-specific concept-based interventions on black-box neural network models, i. e. how the knowledge of high-level attributes may be injected into the network's activations *post hoc*. Inspired by counterfactual explanations, we introduced a simple recipe for editing representations via a probing function trained to predict concept variables at the network's intermediate layer. Such interactions can help end users understand the model through editing and enable efficient human–model collaboration, e. g. in decision support systems.

In addition, we formalised intervention effectiveness as the intervenability measure. In brief, it is quantified by an expected gap in the target prediction loss without and under an intervention. Finally, based on the intervention procedure, we proposed fine-tuning the black-box model by explicitly maximising its intervenability to make the model more reliant on the concept variables without altering its architecture or backbone representations.

In the experiments on the synthetic tabular and natural and medical image data, we observed that black-box models trained without attribute knowledge can, in principle, be intervened on. However, fine-tuning drastically increases intervention effectiveness without impacting pure target or concept prediction performance. From the application perspective, we experimented with deep chest X-ray classifiers and clinical finding prediction and obtained favourable results analogous to those from simpler benchmarks.

7

INTERPLAY BETWEEN EXPLANATION AND FAIRNESS

Previous chapters have focused on interpretable models and explanation techniques with applications to exploratory data analysis and predictive modelling. However, there exist other practical uses of interpretability and explainability, for instance, *model debugging* [41], [131], i. e. detection and removal of biases and problematic patterns in the machine learning model's behaviour. This chapter explores the problem of mitigating, or removing, biases and demonstrates how advances in gradient-based attribution measures (Section 2.4.1) can be leveraged to this end.

One aspect of making machine learning models accountable is ensuring that they are *fair*. This need emerges from an ever-growing demand for ML models in sociotechnical systems [364] and concerns about demographic disparities and discrimination as a result of algorithmic and model-based decision-making [365]. Thus, to ensure transparency of ML models in biomedical and healthcare applications, fairness must be considered alongside the system's explainability and interpretability [366]–[368].

In healthcare applications, ML unfairness and bias often reflect imperfections within the healthcare system and, broadly, society itself. Consequently, these issues may bleed over into algorithmic and high-stakes human decisions informed by ML. Thus, biases require special treatment and mitigation. Specifically, let us consider a concrete example of developing an ICU patient monitoring and management system. If an ML model that is part of the system has been trained on a dataset with few minority group patients, the model might suffer from under- or over-detection of events in these groups. In turn, this increased error rate may lead to alarm fatigue among medical staff and, subsequently, disparate patient treatment and outcomes [369].

A natural research question emerging from the context of this thesis is whether the techniques from interpretable and explainable ML can (i) help detect such harmful behaviours and (ii) mitigate them. Our work tackles the latter question, focusing on deep neural network chest radiograph classifiers, which have been shown to be plagued with biases and spurious correlations [370]–[372].

In the remaining sections, we lay out the basic concepts and background behind algorithmic fairness. We also provide an overview of closely related works on deep chest X-ray classifiers, model pruning, and interpretation of individual units in neural networks. Then, we introduce novel techniques for bias mitigation. We explain the experimental setup and present the results on simple benchmarks and chest X-ray classification tasks. At the end, we discuss and summarise the contributions and results. This chapter is based on the contents and text of the published work "Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods" [373].

7.1 BACKGROUND

To provide a broader context for the novel methods presented later in Section 7.2, we now introduce essential concepts from algorithmic fairness and overview closely related works on bias mitigation, model pruning, the role of individual units in neural networks, and fairness of chest X-ray classifiers.

Throughout this chapter, we will assume being given labelled training, validation, and test data $\mathscr{D} = \{(\boldsymbol{x}_i, a_i, y_i)\}_{i=1}^{N} = \mathscr{D}_{\text{train}} \cup \mathscr{D}_{\text{valid}} \cup \mathscr{D}_{\text{test}}$, where $\boldsymbol{x}_i$ are features, $y_i \in \{0, 1\}$ is the binary label, and $a_i \in \{0, 1\}$ is the so-called *protected attribute* for data point $1 \leq i \leq N$. The protected attribute typically corresponds to the patient's sensitive characteristic, for which a machine learning model may exhibit an unfair behaviour, e. g. age or race. Let $f_{\boldsymbol{\theta}}$ denote a neural network classifier parameterised by $\boldsymbol{\theta}$ and trained on data points $\{(\boldsymbol{x}_i, y_i)\}_i$ from $\mathscr{D}_{\text{train}}$, i. e. without any awareness of the protected attribute values. Generally, we make no specific assumptions on the network architecture. However, in this chapter, we limit the experiments to fully connected and convolutional neural networks. For the fully connected architectures, parameters $\boldsymbol{\theta}$ are given by weight matrices $\left\{\boldsymbol{W}^{\text{in}}, \boldsymbol{W}^{1}, \ldots, \boldsymbol{W}^{L}, \boldsymbol{W}^{\text{out}}\right\}$. For a layer $1 \leq l \leq L$ and input $\boldsymbol{x}$, let $z^{l}(\boldsymbol{x})$ denote preactivations and $h^{l}(\boldsymbol{x}) = \sigma\left(z^{l}(\boldsymbol{x})\right)$ be activations, where σ is a nonlinear activation function. The model's output is then given by sigmoid $\left(\boldsymbol{W}^{\text{out}} h^{L}(\boldsymbol{x})\right)$. The predicted class is $\hat{y} = \mathbf{1}_{\{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \tau\}}$, where $\tau \in [0, 1]$ is a threshold tuned on the validation set. Note that this chapter focuses exclusively on classification and binary protected attributes. However, some discussion within this limited scope might apply to more general output and attribute structures.

7.1.1 *Algorithmic Fairness*

As discussed above, this chapter makes contributions at the interface between explainable ML and *algorithmic fairness*. Fairness is an active research field seeking to identify and correct biases in decision processes based on ML models. As with the interpretability (Section 2.3), there is no one-fits-all definition of fairness in the context of ML. Corbett-Davies *et al.* [374] distinguish two principal approaches: (i) *classification parity* and (ii) *"fairness through unawareness"*. This thesis adopts the former perspective, which we summarily describe below.

Classification parity requires the equality of some form of error rates across the groups of the protected attribute [374]. Two common and practical classification parity measures are *statistical parity* and *equality of opportunity*. Statistical parity difference (SPD) [375], [376] is defined as the difference between the probabilities of the positive prediction across the groups of the protected attribute:

$$\mathrm{SPD} = p\left(\hat{y} = 1 | a = 0\right) - p\left(\hat{y} = 1 | a = 1\right) \tag{7.1}$$

By contrast, the equal opportunity difference (EOD) [375], [377] quantifies the discrepancy in the true positive rates (TPR) of the classifier:

$$\mathrm{EOD} = p\left(\hat{y} = 1 | y = 1, a = 0\right) - p\left(\hat{y} = 1 | y = 1, a = 1\right). \tag{7.2}$$

In practice, quantities from Equations 7.1–7.2 can be estimated by empirical frequencies on a held-out validation set. Of course, it is possible to define more criteria for classification parity building on other types of rates. The ultimate choice among SPD, EOD, and other measures must be informed by the application at hand. For instance, statistical parity may be strongly desirable in some digital marketing applications but impractical in decision support systems for medical diagnosis.

Beyond quantifying the bias using criteria such as the ones above, another important research topic is the design of algorithms to reduce bias, e. g. by directly minimising the criteria from Equation 7.1 or 7.2. Such procedures are known as *debiasing*[1] and typically lead to a tradeoff between the chosen fairness criterion and predictive performance [378]. Intuitively, a debiasing algorithm should reduce bias w.r.t. some measure μ, e. g. given by Equations 7.1–7.2, without sacrificing performance as measured by ρ. Thus,

---

[1] Throughout this chapter, terms "bias" and "debiasing" are not to be confused with their statistical counterparts.

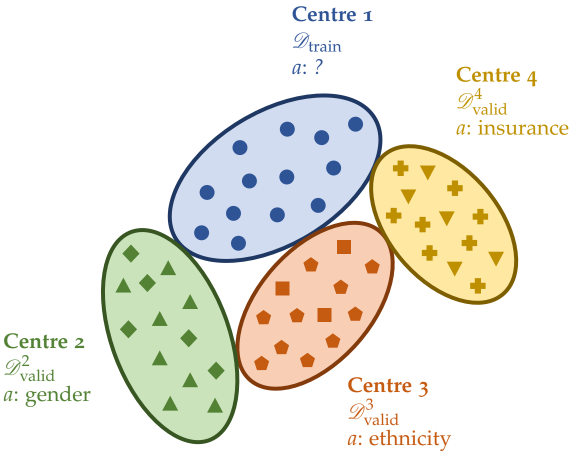

Centre 1
$\mathscr{D}_{\text{train}}$
a: ?
Centre 4
$\mathscr{D}^4_{\text{valid}}$
a: insurance
Centre 2
$\mathscr{D}^2_{\text{valid}}$
a: gender
Centre 3
$\mathscr{D}^3_{\text{valid}}$
a: ethnicity

Figure 7.1: An application scenario for intra-processing. A classifier is trained on the data from centre **1** ($\mathscr{D}_{\text{train}}$) without any regard for the protected attribute (a). Subsequently, the model has to be deployed at centres **2**, **3**, and **4**, which feature local fairness constraints. It is impractical to retrain the model, and instead, we resort to intra-processing and fine-tune the classifier on the local validation sets ($\mathscr{D}^2_{\text{valid}}$, $\mathscr{D}^3_{\text{valid}}$, $\mathscr{D}^4_{\text{valid}}$). Herein, individual points correspond to instances, and marker types denote different protected attribute values.

debiasing amounts to a constrained optimisation problem [375], [379]–[381], where bias is to be minimised subject to the constraints on performance or, *vice versa*, performance is maximised under bias constraints.

Generally, debiasing algorithms may be classified based on the time and manner of their application [375], [382]: (i) preprocessing algorithms are applied *before* training a classifier and usually reweigh or transform training data, obfuscating protected variables and attenuating group disparities [383]–[386], (ii) in-processing techniques incorporate debiasing explicitly into learning using specialised loss functions and regularisers [378], [379], [387], [388], (iii) postprocessing methods treat the model as a black box and agnostically edit its predictions without any effect on the parameters [377], [389], [390], (iv) lastly, intra-processing approaches, first discussed by Savani, White, and Govindarajulu [375], lie in-between in- and post-processing, typically resorting to fine-tuning the model's parameters *post hoc* on a smaller validation set. The chief difference between intra- and postprocessing is that the latter methods family does not require access to the model's parameters but often assumes that the protected attribute is observable at *test time* to adjust the predictions.

This chapter explores the last two method classes. With the widespread availability and use of pretrained models [324], [391]–[393], it may not always be practical to develop a fair classifier from scratch using pre- or in-processing approaches. Moreover, in biomedical and healthcare applications, classifiers might need to be deployed across multiple centres with different *local* fairness considerations and protected attributes of interest. Such application scenarios make intra- and postprocessing techniques especially compelling since the model can be readily fine-tuned or recalibrated on the local data to satisfy newly introduced constraints. Figure 7.1 provides a schematic summary of the application scenario outlined above.

7.1.2 *Biases in Deep Chest X-ray Classifiers*

As mentioned before, an application to chest radiograph classifiers is among the primary subjects of the current chapter. Chest X-ray imaging has become an essential tool for screening and diagnosing conditions affecting the chest and surrounding tissues and organs, requiring specialised training for appropriate interpretation. There have been numerous efforts in developing deep neural network models for screening and computer-aided diagnosis based on chest radiographs [394]–[396] using various datasets [361], [397], [398], with some models achieving a near-expert-level performance [360].

Despite these recent successes, researchers have also scrutinised the fairness of deep classifiers trained on well-established and publicly available chest X-ray datasets [360], [361], [398]. For instance, Larrazabal *et al.* [370] report consistently lower disease prediction AUROCs for underrepresented genders when training on imbalanced datasets. In a multi-centre setting, Zech *et al.* [399] observe that the performance of chest X-ray classifiers is significantly lower on held-out external data, attributing this lack of generalisation to confounding-related biases. Seyyed-Kalantari *et al.* [371], [372] study the underdiagnosis and TPR disparities across three large publicly available chest X-ray datasets, showing higher underdiagnosis and lower TPRs in underserved patient groups.

To the best of our knowledge, few works address the *mitigation* of such biases in computer-aided diagnosis based on chest X-rays. Concurrently and similarly to us, Zhang *et al.* [400] benchmark debiasing techniques for classification parity on publicly available chest radiograph datasets. Zong, Yang, and Hospedales [401] tackle a similar research question for a range of medical image analysis tasks, including X-ray classification.

7.1.3 *Pruning and Individual Units in Neural Networks*

Methodologically, the techniques introduced in Section 7.2 capitalise on the prior literature on model and, specifically, neural network pruning and methods for interpreting the role of individual units within neural networks.

Pruning refers to the procedures for effectively reducing the number of model parameters to induce sparsity (Section 2.3.2), reduce memory and computational complexity, and improve generalisation. Such techniques have been long explored in various model classes, e. g. a classic example is decision trees [402]. In neural networks, pruning usually amounts to removing irrelevant weights or entire structural elements [403], [404], e. g. filters in CNNs. Early approaches, such as *optimal brain damage* [405] and *optimal brain surgeon* [406], leveraged criteria based on the second-order derivatives of the loss function to prune unimportant weights during training. Several modern techniques prune entire structural elements [407]–[409], such as convolutional filters and channels.

In contrast to most attribution measures (Section 2.4.1) elucidating output-input relationships, several recent works investigate the importance and interpretation of *individual* neurons, or units, within deep neural networks. Bau *et al.* [410] study single-unit object detectors whose activations are correlated with high-level concepts in discriminative and generative CNNs. Similarly, Antverg and Belinkov [411] explore the probing of neuron activations in language models. A succession of works [412]–[415] introduce novel attribution measures that quantify the influence of individual neurons.

7.2 PRUNING AND FINE-TUNING FOR DEBIASING NEURAL NETWORKS

This chapter contributes to the line of research on intra-processing debiasing methods [375] (Section 7.1.1, Figure 7.1). In particular, building on previous advances in explaining the role of individual neurons (Section 7.1.3), we introduce criteria and a procedure for debiasing already trained neural networks by pruning individual units. Additionally, we explore the use of these criteria for directly fine-tuning the model's parameters using gradient-based optimisation. In the following sections, we describe differentiable proxy functions for classification parity measures and introduce our pruning and fine-tuning techniques.

7.2.1 *Differentiable Classification Parity Proxies*

As stated in Section 7.1.1, we view algorithmic fairness from the perspective of classification parity, assessing bias using measures such as the SPD and EOD (Equations 7.1–7.2). Observe that, by definition, these classification parity measures are not *differentiable*. Below, we define differentiable proxy functions for the SPD and EOD, facilitating downstream use of gradient-based attribution and optimisation.

Given a dataset $\mathscr{D} = \{(\boldsymbol{x}_i, a_i, y_i)\}_{i=1}^N$ comprising N data points and a classifier $f_{\boldsymbol{\theta}}$, the proxy function $\tilde{\mu}$ for the SPD is given by

$$\tilde{\mu}_{\mathrm{SPD}}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right) = \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\left(1-a_i\right)}{\sum_{i=1}^N 1-a_i} - \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right) a_i}{\sum_{i=1}^N a_i}. \tag{7.3}$$

Analogously, the proxy function for the EOD is

$$\tilde{\mu}_{\mathrm{EOD}}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right) = \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\left(1-a_i\right) y_i}{\sum_{i=1}^N \left(1-a_i\right) y_i} - \frac{\sum_{i=1}^N f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right) a_i y_i}{\sum_{i=1}^N a_i y_i}. \tag{7.4}$$

The proxy functions in Equations 7.3 and 7.4 are similar to the objectives introduced by Zafar *et al.* [379], [380] for fair logistic regression and support vector machine models. In fact, these functions are proportional to the empirical estimators of the (conditional) covariance between the decision boundary of the classifier $f_{\boldsymbol{\theta}}$ and the protected attribute a. We provide a formal derivation of these claims in Appendix E.1, Lemmas E.1.1 and E.1.2.

Intuitively, our methods aim to adjust the parameters of the *already-trained* classifier $f_{\boldsymbol{\theta}}$ using the proxy functions above, specifically by leveraging these proxies as part of pruning criteria or by directly minimising their absolute value via gradient-based learning. Figure 7.2 provides a general summary of the debiasing "protocol" followed by our procedures. This pipeline corresponds to the intra-processing setting described in the previous section (Figure 7.1): firstly, a potentially biased model is developed on the training data, then, it is debiased using our techniques on the validation set, and, finally, the model's performance and bias are assessed on the test data.

Note that, in practice, the choice between the two functions depends on the targeted classification disparity measure, i. e. the SPD or EOD, which, as mentioned, should be chosen based on the application at hand.

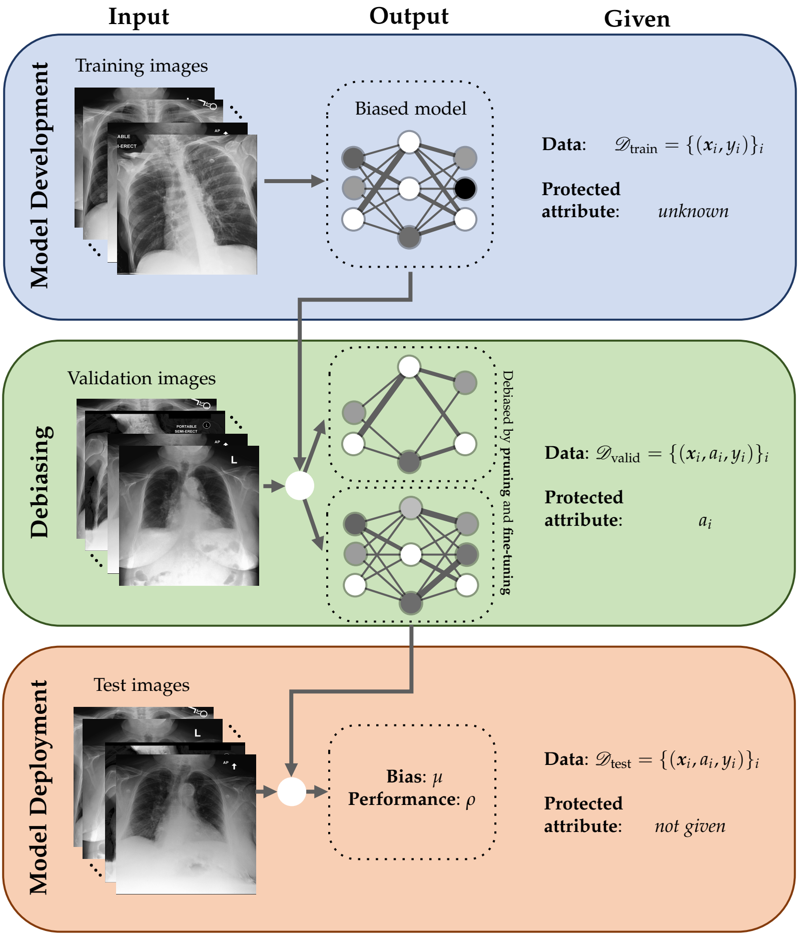

Input
Output
Given
Model Development
Training images
Biased model
Data: $\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_i$
Protected attribute: unknown
Debiasing
Validation images
Debiased by pruning and fine-tuning
Data: $\mathscr{D}_{\text{valid}} = \{(x_i, a_i, y_i)\}_i$
Protected attribute: a_i
Model Deployment
Test images
Bias: μ
Performance: ρ
Data: $\mathscr{D}_{\text{test}} = \{(x_i, a_i, y_i)\}_i$
Protected attribute: not given

FIGURE 7.2: A schematic summary of the debiasing pipeline. **(i)** A model is trained without the knowledge of the protected attribute on $\mathscr{D}_{\text{train}}$. **(ii)** The model is debiased on the validation set ($\mathscr{D}_{\text{valid}}$) using the proposed pruning and fine-tuning procedures. **(iii)** Debiased models are evaluated on the test data ($\mathscr{D}_{\text{test}}$) to assess their predictive performance (ρ) and bias (μ). At prediction, the protected attribute (a_i) is not given to the model and is only utilised for bias evaluation.

7.2.2 *Pruning for Fairness*

In contrast to most previous works [403], [404], which focus on model compression and complexity reduction, we utilise pruning for *bias mitigation*. In the following, we introduce a procedure to prune individual units in a neural network based on their "contribution" to classification disparity.

Building on the influence-directed explanations and internal influence attribution measures proposed by Leino *et al.* [412], we first introduce a gradient-based statistic quantifying an individual unit's contribution to classification disparity, referred to as *gradient-based bias influence*. For a model $f_{\boldsymbol{\theta}}$, differentiable bias proxy function $\tilde{\mu}$, e. g. from Equation 7.3 or 7.4, and the given dataset $\mathscr{D} = \{(\boldsymbol{x}_i, a_i, y_i)\}_{i=1}^{N}$, the gradient-based bias influence of the j-th unit from layer l is given by

$$S_{l,j} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \tilde{\mu}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right)}{\partial z_j^l\left(\boldsymbol{x}_i\right)}, \tag{7.5}$$

where $z_j^l\left(\boldsymbol{x}_i\right)$ denotes the *preactivation* of the j-th unit at input $\boldsymbol{x}_i$. For example, in a fully connected layer, $z_j^l\left(\boldsymbol{x}_i\right)$ is a vector component, whereas, in a convolutional layer, it is a component of a 3D tensor. Regardless of the layer's type and dimensionality, we use a single index j to enumerate the units. Equation 7.5 corresponds to the average value of the partial derivative of the bias proxy function w.r.t. the unit's preactivation. Thus, this measure can help attribute classification disparity to individual neurons.

We utilise the gradient-based bias influence as a criterion to rank and remove the most influential units from the network, effectively reducing classification disparity. Algorithm 4 outlines this pruning procedure comprising a few simple steps. (i) We evaluate the influence $S_{l,j}$ in Equation 7.5 for every unit j in each layer $1 \leq l \leq L$ on the validation data $\mathscr{D}_{\text{valid}}$ (line 3). In practice, the running time and memory complexity of this step can be reduced by evaluating the influence on a subset of $\mathscr{D}_{\text{valid}}$ or by concentrating on a few selected layers. (ii) At each iteration of the procedure, several units are pruned (line 6). The neurons to be removed are chosen according to the criterion $\operatorname{sgn}\left(\mu_0\right) S_{l,j}$, where the gradient-based influence is multiplied by the sign function applied to the bias measure of the unpruned model. The number of units to be pruned per iteration is determined by the number of steps $B \geq 1$. The implementation of pruning is architecture-specific. For

Algorithm 4: Pruning for Debiasing Neural Networks

```
Input: Validation set $\mathscr{D}_{\text{valid}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_i$; trained neural network $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ and $L$ intermediate layers; classification threshold $\tau \in [0,1]$; predictive performance measure $\rho$; bias measure $\mu$; differentiable bias proxy function $\tilde{\mu}$; lower bound on performance $\varrho > 0$; number of steps $B \geq 1$
Output: Pruned network $f_{\tilde{\boldsymbol{\theta}}}$ with parameters $\tilde{\boldsymbol{\theta}}$

Let $\mu_0 \leftarrow \mu\left(\mathbf{1}_{\{f_{\boldsymbol{\theta}}(\cdot)\geq\tau\}}, \mathscr{D}_{\text{valid}}\right)$, where $\mathbf{1}_{\{f_{\boldsymbol{\theta}}(\cdot)\geq\tau\}}$ denotes thresholding applied to the classifier's output
Initialise $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$
Given $\mathscr{D}_{\text{valid}}$ and $\tilde{\mu}$, evaluate $S_{l,j}$ from Equation 7.5 for every unit $j$ in layer $1 \leq l \leq L$
for $b = 0$ to $B - 1$ do
    Let $s_b \leftarrow q_{1-1/B}\left(\left\{\operatorname{sgn}(\mu_0) S_{l,j}\right\}_{l,j}\right)$, where $q_\alpha$ is the empirical $\alpha$-quantile
    Prune unit $j$ in layer $1 \leq l \leq L$ if $\operatorname{sgn}(\mu_0) S_{l,j} > s_b$ and update $\tilde{\boldsymbol{\theta}}$ accordingly, where sgn is the sign function
    Let $\tilde{\tau} \leftarrow \arg\max_{\tau' \in [0,1]} \rho\left(\mathbf{1}_{\{f_{\tilde{\boldsymbol{\theta}}}(\cdot)\geq\tau'\}}, \mathscr{D}_{\text{valid}}\right)$
    Let $\mu_b \leftarrow \mu\left(\mathbf{1}_{\{f_{\tilde{\boldsymbol{\theta}}}(\cdot)\geq\tilde{\tau}\}}, \mathscr{D}_{\text{valid}}\right)$
    Let $\rho_b \leftarrow \rho\left(\mathbf{1}_{\{f_{\tilde{\boldsymbol{\theta}}}(\cdot)\geq\tilde{\tau}\}}, \mathscr{D}_{\text{valid}}\right)$
    Recompute $\left\{S_{l,j}\right\}_{l,j}$ for the pruned network $f_{\tilde{\boldsymbol{\theta}}}$
    Let $\tilde{\boldsymbol{\theta}}_b \leftarrow \tilde{\boldsymbol{\theta}}$
end
Let $b^* \leftarrow \arg\min_{\substack{0 \leq b \leq B-1 \\ \rho_b \geq \varrho}} |\mu_b|$
return $f_{\tilde{\boldsymbol{\theta}}_{b^*}}$
```

fully connected layers, the j-th unit of the l-th layer can be removed by setting relevant weights to zeros: $\boldsymbol{W}^{l}_{j,\cdot} \leftarrow \mathbf{0}$. In convolutional layers, we introduce dropout-like [416] binary masks applied to preactivations $z^{l}\left(\boldsymbol{x}_i\right)$. (iii) The bias μ and predictive performance ρ of the pruned network are assessed on the validation set (lines 7–9). (iv) Lastly, gradient-based bias influence is recomputed for the remaining neurons in the pruned network (line 10), and steps (ii)–(iv) are repeated for several iterations. The final pruned parameter configuration is chosen to minimise the bias on the validation set subject to the lower-bound constraint on predictive performance specified via the parameter $\varrho > 0$ (line 13).

In summary, the procedure in Algorithm 4 greedily removes units from an already-trained neural network classifier to reduce classification disparity. The pruning criterion is a gradient-based attribution measure that quantifies the contribution of individual neurons to a differentiable classification parity proxy function (Section 7.2.1) in the spirit of the influence-directed explanations introduced by the prior literature [412].

7.2.3 *Fine-tuning for Fairness*

While pruning for debiasing may help localise the sources of classification disparity within the network and be combined with model compression, a more straightforward approach to bias mitigation using differentiable proxy functions is their direct gradient-based minimisation.

To this end, we consider fine-tuning the classifier $f_{\boldsymbol{\theta}}$ using the minibatch gradient descent or ascent. Algorithm 5 summarises this procedure in pseudocode. The network's parameters are updated iteratively for E steps using the gradient of the proxy function w.r.t. the model's parameters (line 6). Notably, the direction of the update depends on the initial bias sign. Similar to pruning, the fine-tuned classifier is evaluated on the validation set (lines 7–9), and the parameters minimising the disparity subject to the performance constraint are chosen (line 12).

The fine-tuning procedure has a few additional important hyperparameters. For instance, minibatch size M must be large enough to adequately estimate the terms from Equations 7.3 and 7.4. Similarly, during our experiments, we observed that the learning rate η and the number of fine-tuning steps E should be chosen sufficiently small, not to considerably alter the initial classifier's predictive performance.

Algorithm 5: Fine-tuning for Debiasing Neural Networks

```
Input: Validation set $\mathscr{D}_{\text{valid}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_i$; neural network $f_{\boldsymbol{\theta}}$ with
       parameters $\boldsymbol{\theta}$; classification threshold $\tau \in [0, 1]$; predictive
       performance measure $\rho$; bias measure $\mu$; differentiable bias
       proxy function $\tilde{\mu}$; lower bound on performance $\varrho > 0$; learning
       rate $\eta > 0$; number of steps $E \geq 1$; minibatch size $M \geq 1$
Output: Fine-tuned network $f_{\tilde{\boldsymbol{\theta}}}$ with parameters $\tilde{\boldsymbol{\theta}}$

Let $\mu_0 \leftarrow \mu\left(\mathbf{1}_{\{f_{\boldsymbol{\theta}}(\cdot) \geq \tau\}}, \mathscr{D}_{\text{valid}}\right)$
Initialise $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$
for $e = 0$ to $E - 1$ do
    Draw a minibatch $\mathscr{B} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^{M}$ without replacement, so
      that $\mathscr{B} \subseteq \mathscr{D}_{\text{valid}}$
    Let $\tilde{\mu}_e \leftarrow \tilde{\mu}\left(f_{\tilde{\boldsymbol{\theta}}}, \mathscr{B}\right)$
    Update $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}} - \operatorname{sgn}(\mu_0)\, \eta \nabla_{\tilde{\boldsymbol{\theta}}} \tilde{\mu}_e$
    Let $\tilde{\tau} \leftarrow \arg\max_{\tau' \in [0,1]} \rho\left(\mathbf{1}_{\{f_{\tilde{\boldsymbol{\theta}}}(\cdot) \geq \tau'\}}, \mathscr{D}_{\text{valid}}\right)$
    Let $\mu_e \leftarrow \mu\left(\mathbf{1}_{\{f_{\tilde{\boldsymbol{\theta}}}(\cdot) \geq \tilde{\tau}\}}, \mathscr{D}_{\text{valid}}\right)$
    Let $\rho_e \leftarrow \rho\left(\mathbf{1}_{\{f_{\tilde{\boldsymbol{\theta}}}(\cdot) \geq \tilde{\tau}\}}, \mathscr{D}_{\text{valid}}\right)$
    Let $\tilde{\boldsymbol{\theta}}_e \leftarrow \tilde{\boldsymbol{\theta}}$
end
Let $e^* \leftarrow \arg\min_{\substack{0 \leq e \leq E-1 \\ \rho_e \geq \varrho}} |\mu_e|$
return $f_{\tilde{\boldsymbol{\theta}}_{e^*}}$
```

7.3 EXPERIMENTAL SETUP

This section describes the experimental setup of the chapter. In our experiments, we aim to (i) provide proof-of-concept results on tabular benchmarking datasets supporting the utility of our debiasing methods (Section 7.2) and (ii) explore the use of intra- and postprocessing techniques to mitigate biases in deep chest X-ray classifiers (Section 7.1.2). In the remainder of this section, we introduce the datasets, provide details on the classification models and debiasing methods applied to them, and define evaluation metrics for quantitative comparison.

7.3.1 *Datasets*

We conduct the comparison on tabular and image datasets summarised in Table 7.1. In particular, we include several *nonclinical* benchmarks from the IBM AI Fairness 360 toolkit [382] commonly utilised by the prior methodological literature, for instance, by Savani, White, and Govindarajulu [375]. Below, we briefly describe each dataset; however, a more in-depth discussion of debiasing benchmarks can be found in a thorough survey by Le Quy *et al.* [417].

The first nonclinical tabular benchmark is the Adult Census Income data (**Adult**), containing 48842 instances with seven categorical, two binary, and six numerical features. The classification problem is to predict whether a person's annual income exceeds 50000$ [417], [418]. Our analysis focuses on the protected attribute "sex". Throughout this chapter, we generally refer to protected variables using their original names reported in the data.

Another similar dataset originates from bank phone marketing campaigns (**Bank**) [417], [419]. It comprises 45211 samples with six categorical, four binary, and seven numerical features. The task is to predict a deposit subscription by a potential client. In this case, we consider "age" as the protected variable.

The last nonclinical benchmark we utilise is the Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) dataset [27], [417], consisting of 7214 samples with 31 categorical, 6 binary, and 14 numerical covariates. The underlying problem is the prediction of the risk of recidivism. The conventional protected attribute for this dataset is "race".

Dataset	N_{train}	N_{valid}	N_{test}	p	a	**Architecture**
Adult	27133	9044	9045	98	*Sex*	MLP
Bank	18292	6098	6098	57	*Age*	MLP
COMPAS	3700	1233	1234	401	*Race*	MLP
MIMIC-III	21595	7199	7199	43	*Age*	MLP
	21595	7199	7199	44	*Marital Status*	MLP
	21595	7199	7199	44	*Insurance Type*	MLP
MIMIC-CXR	5528	3368	3426	224×224	*Sex*	CNN
	3984	930	1122	224×224	*Ethnicity*	CNN

TABLE 7.1: A summary of datasets included in the experiments. We report the sizes of the training (N_{train}), validation (N_{valid}), and test (N_{test}) sets, feature dimensionality after preprocessing (p), protected attributes considered (a), and classifier architectures utilised.

As a tabular clinical benchmark, we consider the Medical Information Mart for Intensive Care database (**MIMIC-III**) of patient admissions to critical care units at a large tertiary hospital [420]. We adhere to the preprocessing routine by Purushotham *et al.* [421], retaining only the first admissions of adult patients ($>$ 15 years). Preprocessed data include 17 features from the Simplified Acute Physiology Score (SAPS-II). We average time series variables for every admission and feature. The problem we investigate is the prediction of in-hospital mortality. For the choice and construction of protected attributes, we follow the analysis by Meng *et al.* [422] and focus on "age", "marital status", and "insurance type".

Last but not least, we perform debiasing on the **MIMIC-CXR** dataset of chest radiographs [361] (Section 6.3.1). We retain only frontal view images, resizing them to 224×224 pixels. We focus on "sex" and "ethnicity" as protected variables since the groups of these attributes were previously shown to have disparate classification outcomes [371], [372] (Section 7.1.2). For "sex", we investigate the prediction of "enlarged cardiomediastinum" (enlarged CM), while, for "ethnicity", we choose "pneumonia" as the target variable. In both cases, we utilise studies labelled with "no findings" as the negative class.

7.3.2 *Classification Models and Debiasing Methods*

As reported in Table 7.1, for tabular datasets, we train a fully connected neural network as the classifier to be debiased. The architecture [375] is described in detail in Appendix E.2. For MIMIC-CXR, we study two CNN architectures: VGG-16 [423] and ResNet-18 [339]. All models are trained by minimising the binary CE loss. Additionally, for MIMIC-CXR, we apply random augmentations to images during training to prevent overfitting.

After training the respective classifiers, we perform debiasing on the validation set. Alongside the proposed pruning and fine-tuning procedures (Section 7.2), we evaluate several baseline intra- and postprocessing methods. Specifically, we compare against the closely related intra-processing techniques introduced by Savani, White, and Govindarajulu [375]. The simplest approach is to randomly perturb the parameters of the original classifier using multiplicative Gaussian noise (**Random**) [375]. Perturbation is applied many times, and the parameter configuration maximising the performance subject to bias constraints is returned. Another intra-processing baseline is fine-tuning the classifier using adversarial learning (**Adversarial**) [375] instead of directly minimising the disparity.

From postprocessing algorithms, we apply the reject option classification (**ROC**) [389], which adjusts the classifier's predictions *post hoc* for certain instances falling within a confidence band around the decision boundary. Lastly, we also include the equalised odds postprocessing method (**Eq. Odds**) [377], which adjusts predicted labels probabilistically to balance the odds across protected attribute categories.

7.3.3 *Evaluation Metrics*

As explained above, debiasing is performed exclusively on the validation set. Subsequently, debiased classifiers are evaluated on the test data. Using the MC CV, we train, debias, and evaluate classifiers across several train-validation-test splits. In particular, all models are evaluated w.r.t. their predictive performance and classification disparity. For the former, we report the balanced accuracy. On tabular datasets, for completeness, we report both the SPD and EOD, whereas, on MIMIC-CXR, we focus solely on the EOD since the reduction of the SPD may not be clinically meaningful in the context of medical diagnosis. To explore the model's behaviour throughout pruning and fine-tuning, we visualise changes in performance and bias throughout these procedures.

7.4 RESULTS

This section describes our experimental findings. First, we compare intra- and postprocessing debiasing techniques (Section 7.3.2) on simple tabular benchmarks (Table 7.1) and, subsequently, provide a more in-depth exploration of an application to deep chest X-ray classifiers.

7.4.1 *Benchmarking Results*

We report quantitative results obtained on tabular datasets in Table 7.2. Debiasing is performed separately with the SPD and EOD as disparity criteria. Tables 7.2a and 7.2b contain bias and predictive performance measures, respectively, before and after debiasing. Note that throughout the tables and figures, "Standard" refers to the initial neural network classifier.

Alongside other intra- and postprocessing methods, our pruning and fine-tuning procedures successfully mitigate bias in most datasets (Table 7.2a); overall, our methods tend to sacrifice less accuracy (Table 7.2b). The closely related adversarial intra-processing [375] visibly hurts the performance and, on average, does not reduce the disparity as effectively.

Interestingly, all intra-processing methods considerably reduce the predictive performance when debiasing on the Adult dataset w.r.t. the SPD, attaining results inferior to those of the ROC and equalised odds. Note that the original classifier in this problem has a relatively high SPD. This finding may be attributed to the sensitivity of intra-processing approaches to initial conditions: when the disparity of the classifier is high, intra-processing may reduce the accuracy considerably or fail to mitigate the bias [373].

Additionally, to better understand the effect of pruning and fine-tuning on the neural network, we visualise changes in classification disparity and balanced accuracy throughout these procedures. Figure 7.3 shows the trajectories of the EOD (Figures 7.3b and 7.3d), SPD (Figures 7.3a and 7.3c), and BA obtained on the MIMIC-III data. Observe that both methods successfully drive the bias towards zero while having little effect on the classifier's balanced accuracy. Notably, compared to the training time, relatively few pruning and fine-tuning steps are necessary to mitigate bias. We also observe that in line with the results from Table 7.2, pruning (Figures 7.3a–7.3b) features higher variance in accuracy and bias trajectories than fine-tuning (Figures 7.3c–7.3d). Generally, we have noticed similar behaviour across other tabular datasets; however, we omit these results in the interest of space.

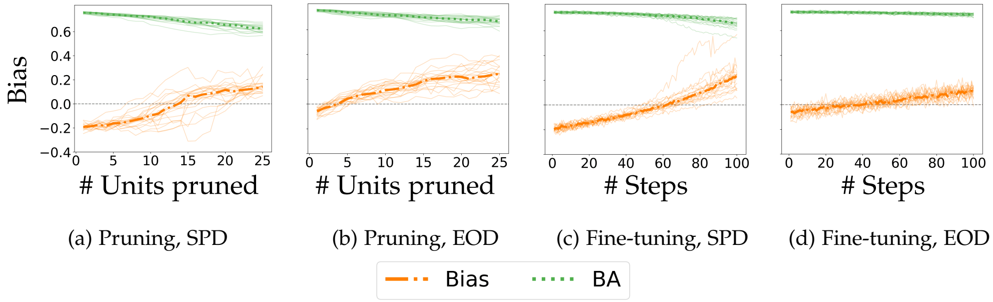
(a) Classification disparity

Bias	**Method**	**Adult:** ***Sex***	**Bank:** ***Age***	**COMPAS:** ***Race***	**MIMIC-III:** ***Age***	**MIMIC-III:** ***Marital***	**MIMIC-III:** ***Insurance***
SPD	Standard	-0.32±0.02	0.18±0.04	0.19±0.03	-0.28±0.03	0.10±0.02	-0.19±0.03
	Random	-0.04±0.01	*0.03±0.04*	0.09±0.04	-0.04±0.01	0.05±0.01	-0.04±0.01
	ROC	-0.04±0.02	0.08±0.04	**-0.01±0.01**	-0.05±0.01	0.03±0.03	-0.05±0.01
	Eq. Odds	-0.09±0.01	0.06±0.03	0.03±0.06	*-0.01±0.01*	*0.01±0.00*	*-0.01±0.00*
	Adversarial	*-0.03±0.00*	0.05±0.03	0.03±0.03	-0.04±0.01	0.04±0.02	-0.03±0.01
	Pruning	-0.04±0.05	**0.02±0.04**	*0.02±0.03*	**0.00±0.02**	**0.00±0.02**	**0.00±0.02**
	Fine-tuning	**-0.01±0.04**	0.04±0.05	0.04±0.04	*-0.01±0.02*	*0.01±0.02*	*-0.01±0.02*
EOD	Standard	-0.14±0.02	0.01±0.04	0.20±0.05	-0.11±0.04	0.08±0.03	-0.05±0.04
	Random	-0.07±0.03	*0.02±0.04*	0.09±0.04	*-0.05±0.05*	0.06±0.04	-0.04±0.04
	ROC	-0.05±0.03	0.04±0.04	**-0.01±0.01**	*-0.05±0.06*	0.03±0.05	-0.04±0.04
	Eq. Odds	**-0.01±0.04**	0.04±0.10	*0.03±0.06*	**0.01±0.04**	**0.01±0.04**	*0.01±0.04*
	Adversarial	-0.09±0.03	0.03±0.06	0.14±0.07	-0.08±0.03	0.06±0.04	-0.02±0.03
	Pruning	**-0.01±0.03**	**0.00±0.07**	0.04±0.06	**0.01±0.06**	*-0.02±0.06*	**0.00±0.04**
	Fine-tuning	*-0.03±0.03*	*0.02±0.06*	0.06±0.06	**-0.01±0.05**	*0.02±0.05*	**0.00±0.04**

(b) Balanced accuracy

Bias	**Method**	**Adult:** ***Sex***	**Bank:** ***Age***	**COMPAS:** ***Race***	**MIMIC-III:** ***Age***	**MIMIC-III:** ***Marital***	**MIMIC-III:** ***Insurance***
SPD	Standard	0.82±0.01	0.86±0.01	0.65±0.01	0.76±0.01	0.76±0.01	0.75±0.01
	Random	0.60±0.01	0.60±0.10	0.60±0.03	0.64±0.01	0.72±0.02	0.67±0.01
	ROC	**0.79±0.01**	0.66±0.10	0.50±0.00	0.63±0.01	**0.75±0.01**	0.68±0.01
	Eq. Odds	*0.73±0.02*	0.70±0.02	0.60±0.01	0.57±0.02	0.57±0.01	0.57±0.01
	Adversarial	0.56±0.01	0.61±0.09	0.56±0.04	0.60±0.02	0.67±0.04	0.64±0.02
	Pruning	0.56±0.04	*0.84±0.01*	*0.63±0.02*	*0.69±0.02*	*0.73±0.01*	*0.69±0.03*
	Fine-tuning	0.66±0.01	**0.86±0.01**	**0.64±0.01**	**0.72±0.01**	**0.75±0.01**	**0.73±0.01**
EOD	Standard	0.82±0.01	0.86±0.01	0.65±0.01	0.76±0.01	0.76±0.01	0.75±0.01
	Random	0.78±0.03	**0.86±0.01**	0.61±0.03	0.72±0.03	*0.74±0.03*	**0.75±0.01**
	ROC	**0.82±0.01**	**0.86±0.01**	0.50±0.00	0.69±0.04	**0.75±0.01**	**0.75±0.02**
	Eq. Odds	0.73±0.02	0.70±0.02	0.60±0.01	0.57±0.02	0.57±0.01	0.57±0.01
	Adversarial	*0.78±0.02*	*0.84±0.01*	0.61±0.02	0.71±0.02	0.73±0.01	0.72±0.03
	Pruning	*0.78±0.02*	**0.86±0.03**	*0.62±0.03*	*0.73±0.01*	0.73±0.02	*0.74±0.01*
	Fine-tuning	**0.82±0.01**	**0.86±0.01**	**0.64±0.01**	**0.75±0.01**	**0.75±0.01**	**0.75±0.01**

TABLE 7.2: (a) Classification disparity and (b) balanced accuracy before and after debiasing neural networks trained on tabular datasets. We report the results w.r.t. statistical parity (SPD) and equal opportunity differences (EOD) separately. Best results are shown in **bold**, *italics* indicates the second best, excluding the original classifier (Standard).



(a) Pruning, SPD (b) Pruning, EOD (c) Fine-tuning, SPD (d) Fine-tuning, EOD

FIGURE 7.3: Changes in the classification disparity, given by the (a, c) SPD and (b, d) EOD, and balanced accuracy of the classifier during (a, b) pruning and (c, d) fine-tuning. The results were obtained on the MIMIC-III dataset by predicting in-hospital mortality with the "insurance type" as the protected attribute. **Bold** lines correspond to the medians across 20 replicates.

7.4.2 *Debiasing Deep Chest X-ray Classifiers*

Finally, we turn to the deep chest X-ray classification experiments on the MIMIC-CXR dataset. As explained in Section 7.3.1, we consider multiple protected attribute and response variable pairs and network architectures.

Table 7.3 summarises the results w.r.t. the EOD and BA. For predicting enlarged CM under the protected attribute "sex", both VGG (Tables 7.3a–Tables 7.3b) and ResNet (Tables 7.3c–Tables 7.3d) exhibit moderate bias, and most methods successfully mitigate it without affecting the BA. Pruning and fine-tuning achieve the best results on average, followed by the equalised odds postprocessing. Similar to the results on tabular data (Table 7.2), adversarial fine-tuning is inferior to our techniques w.r.t. both metrics. The poorer performance may be attributed to overfitting from adversarial training and having to learn discriminator network parameters [375]. In summary, under moderate bias, the proposed techniques reduce classification disparity without a need for retraining from scratch or access to the protected attribute at test time.

By contrast, for predicting pneumonia under the protected attribute "ethnicity", the average EOD of the original model is considerably higher for both architectures (Tables 7.3b and 7.3d). In this case, only the equalised odds postprocessing achieves a satisfactory result. Pruning and fine-tuning do not hurt the predictive performance; however, the average EOD is not

(a) Enlarged CM, *Sex*; VGG-16

Method	**EOD**	**BA**
Standard	-0.05±0.02	0.77±0.01
Random	-0.03±0.03	*0.75±0.01*
ROC	-0.05±0.02	*0.75±0.03*
Eq. Odds	*0.01±0.03*	*0.75±0.01*
Adversarial	-0.04±0.03	0.73±0.01
Pruning	**0.00±0.02**	**0.76±0.02**
Fine-tuning	*-0.01±0.04*	**0.76±0.01**

(b) Pneumonia, *Ethnicity*; VGG-16

Method	**EOD**	**BA**
Standard	-0.14±0.04	0.73±0.02
Random	-0.11±0.06	**0.71±0.02**
ROC	*-0.07±0.06*	0.65±0.06
Eq. Odds	**0.00±0.06**	*0.70±0.01*
Adversarial	-0.13±0.05	*0.70±0.02*
Pruning	-0.09±0.05	**0.71±0.03**
Fine-tuning	-0.08±0.06	**0.71±0.02**

(c) Enlarged CM, *Sex*; ResNet-18

Method	**EOD**	**BA**
Standard	-0.05±0.04	0.76±0.01
Random	**0.00±0.03**	0.73±0.02
ROC	-0.05±0.03	*0.74±0.04*
Eq. Odds	*0.01±0.03*	*0.74±0.01*
Adversarial	-0.04±0.04	0.73±0.02
Pruning	*-0.01±0.03*	*0.74±0.02*
Fine-tuning	**0.00±0.03**	**0.76±0.01**

(d) Pneumonia, *Ethnicity*; ResNet-18

Method	**EOD**	**BA**
Standard	-0.14±0.05	0.73±0.02
Random	*-0.06±0.06*	0.65±0.04
ROC	-0.07±0.04	0.65±0.05
Eq. Odds	**-0.01±0.06**	0.70±0.01
Adversarial	-0.14±0.03	*0.71±0.02*
Pruning	-0.11±0.05	0.70±0.02
Fine-tuning	-0.11±0.05	**0.73±0.02**

TABLE 7.3: Equal opportunity difference (EOD) and balanced accuracy (BA) attained before and after debiasing (a, b) VGG-16 and (c, d) ResNet-18 trained on the MIMIC-CXR to predict (a, c) enlarged cardiomediastinum (CM) with the protected attribute "sex" and (b, d) pneumonia with the protected attribute "ethnicity".

driven to zero. Notably, both methods improve over the naïve random perturbation baseline and adversarial fine-tuning. The poorer performance of intra-processing methods in this instance has several plausible explanations. Specifically, for pneumonia and "ethnicity", the validation set is relatively small (Table 7.1) and may lead to overfitting. Moreover, the protected attribute "ethnicity" is challenging to predict from X-ray images. Therefore, it may be more prudent to resort to postprocessing, which takes the attribute as input at test time. Additionally, note that the attribute "ethnicity" is self-reported [371] and might be noisy and misaligned with the ground truth, introducing another source of variability into our results. Lastly, as mentioned before, intra-processing methods may also be sensitive to the initial bias.

7.5 DISCUSSION

This section contains a discussion of the chapter's contributions and findings in connection to the prior literature and the rest of the thesis. We reflect on methodological and empirical limitations and delve into potential extensions of the current work.

This chapter has focused on the problem of attaining algorithmic fairness [364] from the perspective of classification parity [374]. Alongside interpretability and explainability, the fairness of model-based decisions is arguably among the principal societal considerations in biomedical and healthcare applications of ML. In particular, this chapter provided a comprehensive investigation of intra- and postprocessing debiasing algorithms [375], [377], a practical class of methods that, in contrast to in-processing techniques [379], [380], [388], can be applied *post hoc* to already-trained models, given labelled validation data. From the application perspective, we have concentrated on the deep chest X-ray classifiers, which had been reported to feature disparities across patient subpopulations [371], [372].

To tackle debiasing in the intra-processing setting (Figure 7.1), we introduced novel techniques for pruning and fine-tuning neural networks (Figure 7.2). Building on the works by Zafar *et al.* [379], [380], we considered differentiable proxy functions for the SPD and EOD (Equations 7.3 and 7.4), showing their relation to the covariance between the classifier's decision boundary and protected attribute values (Appendix E.1).

We utilised these proxy functions to formulate a gradient-based pruning criterion (Equation 7.5), quantifying the influence of individual units within a neural network on classification disparity. This criterion is closely related to gradient-based attribution measures (Section 2.4.1), particularly influence-directed explanations by Leino *et al.* [412]. Thus, we leveraged previous efforts at understanding and explaining the role of individual neurons (Section 7.1.3) to perform structured pruning for bias mitigation. Similar to the technique described in Chapter 6, our pruning procedure (Section 7.2.2, Algorithm 4) capitalises on the class of *post hoc* explanation methods, albeit with a different purpose—reducing bias.

In addition to pruning, we introduced a fine-tuning procedure for debiasing (Section 7.2.3, Algorithm 5), which directly minimises bias proxy functions using minibatch gradient descent. Since our proxy functions are differentiable, this method, unlike related in-processing [388] and fine-tuning approaches [375], does not require unwieldy adversarial learning and the introduction of additional parameters.

In the broader context of the literature, this chapter contributes to the research on bias mitigation in classification models. While our methods share similarities with the prior literature, below, we discuss several distinctive features. Our differentiable proxy functions are closely related to the objectives considered by Zafar *et al.* [379], [380]. However, their experimental scope is limited to linear and kernel-based classification, whereas we focus primarily on neural networks.

Although methods for removing biases from neural networks are abundant [375], [378], [388], [424], most of these works concentrate on the in-processing setting, where protected attributes are part of the *training* set, and resort to *adversarial* learning. By contrast, we investigate the intra-processing scenario and propose simple and effective criteria that can be used for network pruning or minimised *directly* without training a discriminator network. The most closely related technique is adversarial fine-tuning proposed by Savani, White, and Govindarajulu [375], who similarly study intra-processing methods. However, in practice, we observed that their approach is prone to overfitting and does not attain empirical performance comparable to ours (Section 7.4). Moreover, other algorithms introduced by these authors are poorly scalable, relying on computationally expensive zeroth-order optimisation.

Model-agnostic postprocessing methods, such as ROC [389] and equalised odds [377], comprise a compelling alternative to intra-processing. We observed that these techniques performed relatively well throughout most experiments. Nevertheless, their significant drawback is that the protected attribute must be observable at *test* time, which, in contrast, is not assumed under the intra-processing scenario.

Next to developing novel debiasing algorithms, this chapter contributes empirically to the research on biases in deep chest radiograph classification models (Section 7.1.2). The methodological works discussed above predominantly explore tabular and natural image benchmarks, not delving into biomedical and healthcare applications. Moreover, previous literature has identified biases within the state-of-the-art models trained on large-scale publicly available datasets [370]–[372]. However, these works have not investigated the *mitigation* of such biases. Thus, to the best of our knowledge, ours and the concurrent efforts by Zhang *et al.* [400] and Zong, Yang, and Hospedales [401] are among the first comprehensive empirical studies in this specific.

In particular, our experiments have explored tabular and medical image datasets and fully connected and convolutional neural networks. In tabular benchmarks, proposed intra-processing approaches effectively reduce the bias and offer improved predictive performance over model-agnostic postprocessing techniques (Table 7.2). We have also demonstrated the utility of our procedures on the MIMIC-CXR dataset for VGG-16 and ResNet-18 architectures (Table 7.3). While effective in many settings, introduced pruning and fine-tuning methods exhibited poorer results when the validation set was too small and the initial bias of the classifier was too high. In such cases, it may be prudent to redevelop the model from scratch or resort to postprocessing.

7.5.1 *Limitations*

Let us now reflect on the limitations of our methods and experimental setup. To begin with, this chapter provides a limited perspective on algorithmic fairness. Specifically, we have restricted our discussion to *classification parity* (Section 7.1.1). However, alternative technical definitions of bias have been explored by the literature [374]. Consideration of other frameworks, for instance, based on calibration [390], would facilitate a more comprehensive treatment of the subject. Moreover, our analysis has exclusively focused on *binary* classification problems under *binary* protected attributes with the SPD and EOD as disparity measures. To accommodate more complex modelling tasks, it would be necessary to adapt our methods to multilevel protected and response variables and consider other definitions of disparity, e. g. explored by Zafar *et al.* [380].

On a higher level, the intra-processing setup (Figure 7.2) we investigate has practical restrictions. Although intra-processing algorithms do not require access to the protected attribute at training and test time, they still need a labelled validation set to adjust the classifier's parameters. When debugging an ML model, a practitioner may be altogether unaware of potential sources of bias. Therefore, the discovery and mitigation of biases in the absence of protected attribute labels is a relevant and open research question. Previous attempts at tackling this more challenging task impose additional assumptions on the nature of the classification problem and utilise representation learning [425] and generative modelling approaches [426].

Beyond the methodological scope, our experimental setup likewise has several limitations. To provide proof-of-concept results, some datasets had to be simplified. For instance, similar to Meng *et al.* [422], we investigated the insurance type as a protected attribute in the MIMIC-III dataset (Table 7.1), grouping Medicare and Medicaid insurance types into a single category despite these being substantially different programmes. Similarly, for the MIMIC-CXR, we considered a smaller subset of radiographs to reduce training time, developing our models on frontal-view images and only including X-rays without *any* findings into the negative class. Thus, it would be interesting to explore a more realistic scenario with multiple disease classes and views while also considering additional protected attribute and label pairings.

Another noteworthy limitation is that we evaluate and compare the methods on three neural network architectures. The extension of our findings to other CNN model designs [427], [428] and architecture classes may require nontrivial adjustments, especially in the pruning procedure.

7.5.2 *Future Work*

Beyond addressing the limitations and open questions outlined above, this work opens many other promising directions for future research. Our pruning procedure relies on *gradient-based* attribution to remove units in the *intermediate* layers of a neural network. Firstly, alternative pruning criteria, which do not require the costly computation of partial derivatives, should be explored. Secondly, with some adjustment, pruning could be utilised as an input feature selection technique to remove unwanted covariates from the predictive model. The latter could prove instrumental in tabular datasets, where features are high-level and human-understandable. As explained in Section 7.1.3, model compression is among the primary goals of pruning neural networks [403]. Thus, future work could combine model compression techniques with our debiasing method to incorporate fairness considerations, especially given the empirical evidence that naïve pruning strategies can adversely affect fairness [429]. Another question underexplored by this chapter is the localisation of disparity within the neural network's architecture. Since the gradient-based influence statistic we introduce quantifies the contribution of individual neurons to the disparity, we can readily visualise the "location" of bias in the network to, for instance, compare and understand the role of layers and modules in the final output. Such analysis can improve the *mechanistic* interpretability [430] of neural networks, specifically in regard to algorithmic fairness.

7.6 SUMMARY

This chapter has described methods at the interface between algorithmic fairness and *post hoc* explainability. We introduced differentiable functions that serve as proxies for classification disparity and utilised them as the pruning criteria and objectives for neural network debiasing. In particular, we investigated debiasing under the intra-processing scenario, where the model's parameters are adjusted *post hoc* to incorporate fairness constraints on a smaller validation dataset.

The pruning criterion we proposed is inspired by gradient-based attribution measures for individual neural network units. Thus, it allows removing neurons with the highest contribution to the classification disparity. In a similar vein, we also considered fine-tuning the classifier by directly minimising a differentiable bias proxy. Both techniques form the core methodological contribution of this chapter, complementing prior literature on intra-processing algorithms.

Furthermore, we conducted comprehensive experiments to compare intra- and postprocessing debiasing methods on tabular benchmarks and the MIMIC-CXR dataset comprising chest radiograph images. Previous works have reported numerous biases in deep chest X-ray classification. Our empirical findings in this regard are among the first efforts in exploring the mitigation of such biases. In general, we observed favourable results suggesting that black-box neural networks can be debiased *post hoc* by pruning or fine-tuning. In many cases, our algorithms outperformed closely related adversarial-learning-based approaches.

8

CONCLUDING REMARKS

This dissertation has introduced several interpretable and explainable ML models and methods tailored towards biomedical and healthcare data analysis problems. Following a data-, application-, and user-driven perspective, we have demonstrated that, beyond social and ethical value, interpretability and explainability help in (i) performing exploratory data analysis (Chapters 3–4), (ii) supporting medical professionals' decisions (Chapter 5), (iii) facilitating interaction with the users (Chapter 6), and (iv) model debugging (Chapter 7). In this final chapter, we make concluding remarks and recapitulate our contributions in a broader context. Additionally, we reflect on the general limitations of our techniques and experimental setups and discuss future research directions.

In the *Introduction* to this thesis (Chapter 1), we have posed two research questions (Questions 1 and 2), which we have subsequently addressed in Parts I and II. In particular, our goals were to (1) develop and incorporate inductive biases to render neural network models interpretable (*ante hoc*) and (2) leverage explanation methods to interact with and edit already-trained black-box models (*post hoc*). The novel techniques described in this work span several model classes and method families (Table 1.1) and are motivated by classic problems arising in biomedical and healthcare domains, such as time series, survival, and medical image analysis (Figure 1.1). Thus, beyond methodological contributions, another focal point of our research is the empirical evaluation on *realistic* biomedical and healthcare benchmarks.

8.1 SUMMARY OF CONTRIBUTIONS

We began this thesis by providing a scoping review of the recent literature in Chapter 2, covering both *ante hoc* interpretable (Section 2.3) and *post hoc* explanation (Section 2.4) approaches. By contrast to similar literature surveys [25], [41], [147], this chapter went beyond high-level notions and closely examined *concrete* model class and method family examples. In addition to the methods-centric perspective, we have briefly touched on the specifics of interpretability and explainability in biomedicine and healthcare applications in Section 2.5. Given the diversity of the overviewed techniques, an essential takeaway from this chapter reflected throughout the rest of our work is that, to date, the literature has produced no (sufficiently specific) "golden-bullet" definition for an interpretable model or an explanation. In our view, one reason for the lack of plausible definitions is that interpretability and explainability must be application- and user-specific.

With Chapter 3, we turned to the first of the two parts of this dissertation, contributing novel models and methods [153], [214], [292], [352], [373]. We investigated the utility of interpretable models for *exploratory data analysis*, concretely, designing models that can yield valuable insights into the relations between the observed variables (*structure learning*). In particular, this chapter tackled time series analysis [30] from the perspective of Granger-causal inference [152]. Combining aspects of self-explaining neural networks (Section 2.3.4) [84], varying-coefficient models [81] (Section 2.3.3), and vector autoregression [30], we introduced a neural-network-based model that, as shown empirically, can accurately infer and interpret temporal relationships among longitudinally observed variables. We also demonstrated that the proposed architecture is more accurate *at inference* than related interpretable models based on the attention mechanism [101] or sparse-input neural networks [95], [96] (Section 2.3.4).

Continuing on exploratory data analysis, Chapter 4 investigated a different approach to interpretability, wherein nonlinear relationships are explained using *prototypes* (Section 2.4.4) defined in the feature space. We tailored our method to the problem of *survival analysis* [207], routinely arising in the healthcare domain. Capitalising on variational autoencoders [105] for clustering [235] and deep survival analysis [209], we developed a probabilistic generative model well-suited for unstructured and high-dimensional data types, such as medical images. A latent-space finite mixture of regression models trained jointly with a VAE allows the discovery of patient subgroups driven by *both* the covariates and time to event. Such clusters can help practitioners visualise and understand heterogeneities in the marginal distributions of features and survival time and variability in their conditional relationship. Furthermore, mixture modelling enables prototype-based explanation of the relationship between the features and response, where the clusters and their centroids serve as prototypes. This approach to interpretability resembles the Bayesian case model [138], which exploits clustering with a similar goal. A noteworthy contribution of this chapter is the application of our model and similar neural-network-based approaches to a challenging dataset of computed tomography scans from NSCLC patients. Our experiments showed that the proposed model discovers subgroups with more pronounced phenotypic differences than those uncovered using alternative approaches [213], [228], highlighting essential associations between the covariates and survival time.

Next, Chapter 5 turned to a higher-level perspective on interpretability. In contrast to the previous chapters, where interpretations primarily operated in the raw feature space, here, we considered a *concept-based* approach (Section 2.4.3) reliant on parsimonious high-level attributes. Concretely, we studied concept bottleneck models [22] in medical *image interpretation* and clinical *decision support*. Motivated by the properties pertinent to medical imaging datasets, we introduced enhancements to CBMs. Firstly, we extended the model to the *multiview learning* scenario [32] by fusing view-specific representations to predict concept variables. Secondly, we tackled the problem of *unobserved concepts* by introducing an additional branch to learn disentangled complementary representations. The primary application studied in this chapter was the prediction of the diagnosis, treatment assignment, and disease severity in children with suspected appendicitis [305] based on multiple views from abdominal ultrasonography [323]. Our results showed that the proposed enhancements successfully scale CBMs to more challenging classification tasks, making them a viable alternative to black-box architectures.

Taken together, Chapters 3–5 proposed various *interpretable* neural-network-based models to solve problems routinely arising in biomedical and healthcare data analysis. Addressing Question 1 raised in Chapter 1, we explored self-explaining neural networks, deep latent variable models paired with mixture regression, and concept bottleneck models as alternatives to black-box architectures. In real-world applications, we demonstrated that our models provide valuable insights into relations between the covariates and responses, arguably the main desired output of interpretable ML [10].

In the second part of this dissertation (Chapters 6–7), we shifted our attention to *post hoc* explanations (Section 2.4). In particular, we addressed Question 2, investigating how explanation techniques may be leveraged to interact with and edit *pretrained* neural networks. Similar to Part I, our primary consideration was biomedical and healthcare application scenarios.

Thus, Chapter 6, similar to the previous one, explored concept-based methods but in application to already-trained models. A compelling feature of CBMs [22] is the user's ability to steer their output via concepts by so-called interventions. To this end, we proposed a simple procedure for *concept-based interventions* on an intermediate layer of a black-box neural network. Via a probing function [293], [294] trained to predict attributes, our method edits the network's activations to align them more closely with the given concept values. In addition, we formalised a measure of the effectiveness of such interventions and introduced a procedure to explicitly fine-tune black box models to increase intervention impact. Our experiments included the evaluation on publicly available chest radiograph datasets [360], [361], suggesting that our fine-tuning procedure results in significantly more effective interventions than CBMs trained naïvely *post hoc* on the network's activations [343], [351].

Subsequently, Chapter 7 also tackled *model editing* [363] but with a different purpose: to mitigate biases in neural network classifiers [388]. *Algorithmic fairness* [364] remains an essential societal consideration and is equally pertinent to biomedical and healthcare domains. Given emerging evidence for single-neuron object detectors in neural networks [410], this chapter investigated the use of gradient-based attribution explanations (Section 2.4.1) to *prune* [403] individual units contributing to classification disparity [376], [377]. We formulated differentiable proxy functions for two common disparity measures and, leveraging these and gradient-based attribution, defined pruning criteria. Additionally, we investigated the utility of these proxies to directly fine-tune neural network models. Our experiments mainly concentrated on debiasing deep chest X-ray classifiers [361], which had been previously shown to be biased w.r.t. several sensitive attributes [371], [372], such as gender and ethnicity. We demonstrated that our techniques effectively reduce disparities without considerably affecting the predictive performance.

In summary, Part II of this thesis treated *post hoc* explanation techniques for pretrained neural networks. It explored how such methods can be tailored to and leveraged for *human–model interaction* and *model editing* in the context of medical imaging data. We showed that, beyond mere introspection, explainable ML can help make predictive models fairer and more user-friendly.

8.2 LIMITATIONS AND OUTLOOK

Both interpretable and explainable ML and biomedical and healthcare applications research pose many challenges and open questions either at the periphery or beyond the scope of the current dissertation. Below, we will reflect on the general recurring limitations of our work and promising directions for future exploration. For the discussion of more specific points, we refer interested readers to the individual chapters.

The distinction between *ante hoc* interpretable models and *post hoc* explanation methods [11] that we explained in Chapter 2 proved instrumental in understanding the application scope of various techniques. However, these two, by now, conventional paradigms must not limit future research. For instance, Madsen *et al.* [431] outline promising alternatives that could resolve some limitations of the *ante* and *post hoc* techniques. Among them are *self-explanations* produced by language models. While, at the time of writing, it is not definite if these models can provide *faithful* explanations, enforcing their faithfulness and consistency is an open challenge that, if addressed, would improve the interpretability of the state-of-the-art generative approaches. In a similar vein, incorporating other modalities, such as natural language [432], e. g. through VLMs, may help generate more *interactive* and *intuitive* interpretations and explanations.

Another noteworthy perspective not explored in this thesis, yet relevant to a more fundamental understanding of neural network models, is *mechanistic interpretability* [433]. In contrast to the approaches we concentrated on that rely on prototypes or simple attributes, this research direction tries to establish a lower-level understanding of the features, neurons, and intricate circuitry embedded in modern architectures, such as transformers [98]. Such analyses can produce more fine-grained and faithful explanations of the model's behaviour.

Our experiments primarily focused on proof-of-concept findings and smaller neural network architectures. Given the moderate dataset sizes in the current biomedical and healthcare research, the use of larger and more complex architectures may not always be justifiable. Nevertheless, with the spotlight shifting towards *foundational models* [8], [393] that are highly reusable and performant across many tasks, interpretable and explainable ML should explore this new frontier.

Some chapters of our works are situated at the interface between interpretability and explainability and other technical problems. For example, in Chapter 5, we have briefly touched on multiview and multimodal learning [32], [33], and Chapter 7 is related to the topic of algorithmic fairness [364]. Generally, many more connections to other subfields of machine learning and statistics could be drawn. In particular, interpretable and explainable ML could tap into other perspectives, e. g. probabilistic modelling [240], uncertainty estimation [434], causality [14], and domain generalisation [435]. We foresee both (i) ideas from these subfields helping the design of better-informed interpretable models and explanation techniques, and

(ii) interpretations and explanations being leveraged to address problems relevant to these external areas of interest.

Another limitation of the many model and method families we studied is their reliance on *strong supervision* [38] (Table 1.1), in the sense that they assume access to additional attribute annotations in the training or validation set. While in practice, biomedical and healthcare datasets are often densely annotated, and we should ideally leverage annotations observed alongside the response variable, strong supervision remains a very restrictive learning scenario due to the costs associated with annotation. Thus, future research should focus on the automated discovery of high-level attributes, e. g. as preliminarily explored in some related works leveraging multimodal approaches [343], [351]. Furthermore, similar problems are central to the representation learning techniques that aim to disentangle and identify generative factors [285], [436], typically in a weakly-supervised fashion.

Although inspired by biomedical and healthcare applications, much of our discussion has focused on technical aspects. Thus, application-grounded [15] and open-ended evaluation of our models and methods has been left for future work. In our view, the full utility of interpretable and explainable ML for different stakeholders in healthcare [34] can only be truly evaluated through comprehensive user studies in direct controlled comparison with alternative black-box approaches. Needless to mention, for practical reasons, such evaluation is hard to implement, especially in a setting close to clinical.

Lastly, while this dissertation has treated diverse data types and tasks (Table 1.1), biomedical and healthcare domains feature many other complex input and output structures. Thus, future efforts should be invested in developing interpretable models and explanation techniques for different modalities, e. g. graphs [437] utilised in drug design or videos [438] prevalent in computer-assisted surgery. Moreover, the list of tasks we investigated is not exhaustive, and interpretable and explainable ML can be relevant in many other scenarios, such as anomaly detection or recommendation.

A

SUPPLEMENTARY: NONLINEAR TIME SERIES STRUCTURE LEARNING

A.1 NETWORK ARCHITECTURES

As explained in Section 3.2, the GVAR model utilises neural networks $\left\{\mathbf{\Phi}_{\boldsymbol{\theta}_k}\right\}_{k=1}^{K}$ to map time series values to generalised coefficient matrices. Table A.1 contains pseudocode describing the architecture utilised for these networks across most of the experiments discussed in Section 3.3. `Linear(`m`)` corresponds to a fully connected layer with m output units, and `ReLU()` denotes the rectified linear unit activation function.

	$\mathbf{\Phi}_{\boldsymbol{\theta}_k}$
1	`Linear(50);ReLU()`
2	`Linear(50);ReLU()`
3	`Linear(50);ReLU()`
4	`Linear(`p^2`)`

Table A.1: Neural network architecture used to map lagged time series values to generalised coefficient matrices $\left\{\mathbf{\Phi}_{\boldsymbol{\theta}_k}\right\}_{k=1}^{K}$ (Equation 3.6). Herein, p denotes the number of time series variables.

B

SUPPLEMENTARY: PROTOTYPE-BASED EXPLANATIONS FOR DEEP SURVIVAL ANALYSIS

B.1 VARIATIONAL AUTOENCODERS: ELBO DERIVATION

Following the setting and notation outlined in Section 4.1.2, for VAEs, the evidence lower bound can be derived as follows:

$$\begin{aligned}
\log p_{\theta}(\boldsymbol{x}) &= \log \int p_{\theta}(\boldsymbol{x}|\boldsymbol{z})\, p_{\theta}(\boldsymbol{z})\, d\boldsymbol{z} && \text{(B.1)}\\
&= \log \int p_{\theta}(\boldsymbol{x}|\boldsymbol{z})\, p_{\theta}(\boldsymbol{z}) \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} && \text{(B.2)}\\
&= \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \frac{p_{\theta}(\boldsymbol{x}|\boldsymbol{z})\, p_{\theta}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} && \text{(B.3)}\\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \log \frac{p_{\theta}(\boldsymbol{x}|\boldsymbol{z})\, p_{\theta}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} && \text{(B.4)}\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \log p_{\theta}(\boldsymbol{x}|\boldsymbol{z}) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \log \frac{p_{\theta}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} && \text{(B.5)}\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \log p_{\theta}(\boldsymbol{x}|\boldsymbol{z}) - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,\|\, p_{\theta}(\boldsymbol{z})\right), && \text{(B.6)}
\end{aligned}$$

where Equation B.4 follows from the Jensen's inequality.

B.2 SYNTHETIC DATA GENERATION

As explained in Section 4.3.1, the generative procedure for nonlinear synthetic data is reminiscent of the process assumed by the VaDeSC and consists of the following steps:

1. Initialise prior cluster assignment probabilities $\pi_j = 1/K$ for $1 \leq j \leq K$.
2. Sample cluster assignments $c_i \sim \mathrm{Categorical}(\boldsymbol{\pi})$ for $1 \leq i \leq N$.
3. Generate cluster-specific mean vectors $\boldsymbol{\mu}_i \in \mathbb{R}^J$ for $1 \leq i \leq K$ by sampling $\mu_{i,j} \sim \mathrm{Uniform}(-0.5, 0.5)$ for $1 \leq j \leq J$.
4. Initialise cluster-specific covariance matrices $\boldsymbol{\Sigma}_i = \mathrm{diag}(\boldsymbol{s}_i)$ for $1 \leq i \leq K$, where $\boldsymbol{s}_i \in \mathbb{R}^J$ are randomly-generated vectors with positive components.

5. Sample representations $\boldsymbol{z}_i \sim \mathcal{N}\left(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}\right)$ for $1 \leq i \leq N$.
6. Initialise $g\left(\boldsymbol{z}\right) = \boldsymbol{W}_2\mathrm{ReLU}\left(\boldsymbol{W}_1\mathrm{ReLU}\left(\boldsymbol{W}_0\boldsymbol{z} + \boldsymbol{b}_0\right) + \boldsymbol{b}_1\right) + \boldsymbol{b}_2$, where $\boldsymbol{W}_0 \in \mathbb{R}^{h\times J}$, $\boldsymbol{W}_1 \in \mathbb{R}^{h\times h}$, $\boldsymbol{W}_2 \in \mathbb{R}^{p\times h}$ and $\boldsymbol{b}_0, \boldsymbol{b}_1 \in \mathbb{R}^{h}$, $\boldsymbol{b}_2 \in \mathbb{R}^{p}$ are randomly-generated matrices and vectors. Herein, ReLU denotes the rectified linear unit activation function applied elementwise.
7. Let $\boldsymbol{x}_i = g\left(\boldsymbol{z}_i\right)$ for $1 \leq i \leq N$.
8. Generate cluster-specific coefficient vectors $\boldsymbol{\beta}_i$ for $1 \leq i \leq K$ by sampling $\beta_{i,j} \sim \mathrm{Uniform}\left(-10, 10\right)$ for $1 \leq j \leq J$.
9. Sample uncensored survival times $u_i \sim \mathrm{Weibull}\left(\mathrm{softplus}\left(\boldsymbol{\beta}_{c_i}^{\top}\boldsymbol{z}_i\right), k\right)$ for $1 \leq i \leq N$, where k is the prespecified shape parameter.
10. Sample censoring indicators $\delta_i \sim \mathrm{Bernoulli}\left(1 - p_{\mathrm{cens}}\right)$ for $1 \leq i \leq N$, where $0 \leq p_{\mathrm{cens}} < 1$ is the prespecified probability of censoring.
11. Sample $\tilde{u}_i \sim \mathrm{Uniform}\left(0, u_i\right)$ for $1 \leq i \leq N$ and define observed survival times as $t_i = \delta_i u_i + (1 - \delta_i)\tilde{u}_i$.

In the experiments outlined in Section 4.3, we generate $N = 60000$ data points with $p = 1000$ features originating from $K = 3$ clusters under the latent space dimensionality of $J = 16$. The survival times are sampled from the Weibull distribution with the shape parameter of $k = 1$ with the probability of censoring $p_{\mathrm{cens}} = 0.3$.

Another synthetic dataset introduced in Section 4.3.1 is the survMNIST. The underlying generative procedure closely follows the original implementation in [251]:

1. Assign each digit in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ to one of K clusters s.t. every cluster includes at least one digit.
2. Based on the digits' assignment to clusters, initialise cluster assignments c_i for data points $1 \leq i \leq N$.
3. Generate cluster-specific risk scores by sampling $r_i \sim \mathrm{Uniform}\left(0.5, 15\right)$ for $1 \leq i \leq K$.
4. Initialise cluster-specific scale parameters as $\lambda_i = \frac{1}{t_0}\exp\left\{r_i\right\}$ for $1 \leq i \leq K$, where t_0 is the prespecified mean survival time.

5. Generate the uncensored survival times $u_i = -\frac{\log a_i}{\lambda_{c_i}}$ by sampling $a_i \sim \text{Uniform}\,(0, 1)$ for $1 \leq i \leq N$.

6. Compute the empirical quantile $q_{\text{cens}} = q_{1-p_{\text{cens}}}\left(\{u_i\}_{i=1}^{N}\right)$, where q_α denotes the empirical α-quantile and p_{cens} is the prespecified *lower bound* on the probability of censoring.

7. Sample the global censoring time $\tilde{u} \sim \text{Uniform}\,(\min_{1\leq i\leq N} u_i, q_{\text{cens}})$.

8. Let the censoring indicators be given by $\delta_i = \mathbf{1}_{\{u_i \leq \tilde{u}\}}$ for $1 \leq i \leq N$.

9. Define observed survival times as $t_i = \delta_i u_i + (1 - \delta_i)\tilde{u}$ for $1 \leq i \leq N$.

Observe that, in contrast to the generative procedure for the synthetic tabular data provided at the beginning of this appendix, p_{cens} is only a lower bound on the percentage of censored survival times. In all experiments, we set $p_{\text{cens}} = 0.3$.

B.3 NETWORK ARCHITECTURES

Tables B.1–B.3 provide pseudocode descriptions of the encoder and decoder neural network architectures utilised for the VaDeSC and baseline models (Section 4.3). The notation is similar to the one introduced in Appendix A.1. In Table B.3, `Conv2D(`C`, kernel_size)` is a 2D convolutional layer with C output channels and a kernel size given by `kernel_size`; `MaxPool2D(kernel_size)` is a 2D max pooling operation specified by the given kernel size; `Upsampling2D(scale_factor)` is a 2D upsampling operation with the given scaling factor; and `ConvTranspose2D(`C`, kernel_size)` is a transposed 2D convolutional layer.

	Encoder
1	`Linear(500); ReLU()`
2	`Linear(500); ReLU()`
3	`Linear(2000); ReLU()`
4	`mu = Linear(`J`); sigma = Linear(`J`)`

	Decoder
1	`Linear(2000); ReLU()`
2	`Linear(500); ReLU()`
3	`Linear(500); ReLU()`
4	`Linear(`p`)`

TABLE B.1: Encoder and decoder neural network architectures for the synthetic, survMNIST, and HGG data. Herein, J corresponds to the number of the latent space dimensions, and p is the number of input features. Depending on the data type, a relevant activation function is applied to the decoder's output, e. g. the sigmoid function for the survMNIST.

	Encoder
1	`Linear(50); ReLU()`
2	`Linear(100); ReLU()`
3	`mu = Linear(`J`); sigma = Linear(`J`)`

	Decoder
1	`Linear(100); ReLU()`
2	`Linear(50); ReLU()`
3	`Linear(`p`)`

TABLE B.2: Encoder and decoder neural network architectures for the SUPPORT, FLChain, and Hemodialysis data.

	Encoder
1	`Conv2D(32, 3);ReLU()`
2	`Conv2D(32, 3);ReLU()`
3	`MaxPool2D(2)`
4	`Conv2D(64, 3);ReLU()`
5	`Conv2D(64, 3);ReLU()`
6	`MaxPool2D(2);Flatten()`
7	`mu = Linear(`J`);sigma = Linear(`J`)`

	Decoder
1	`Linear(10816);Reshape((13, 13, 64))`
3	`Upsampling2D(2)`
4	`ConvTranspose2D(64, 3);ReLU()`
5	`ConvTranspose2D(64, 3);ReLU()`
6	`Upsampling2D(2)`
7	`ConvTranspose2D(32, 3);ReLU()`
8	`ConvTranspose2D(32, 3);ReLU()`
9	`ConvTranspose2D(1, 3)`

TABLE B.3: Encoder and decoder neural network architectures for the NSCLC data. Architectures include (de)convolutional blocks from the VGG network proposed by Simonyan and Zisserman [423].

C

SUPPLEMENTARY: CONCEPT-BASED MODELS IN THE WILD

C.1 SYNTHETIC DATA GENERATION

The synthetic nonlinear dataset briefly described in Section 5.3.1 features nonlinear relationships between the covariates, concepts, and labels. In particular, its generative procedure closely resembles the forward pass of the CBM [22]. Following the notation introduced in Section 5.1, let N, p, V, and K denote the number of data points, covariates *per view*, views, and concepts, respectively. The data-generating process comprises the following steps:

1. Randomly draw a vector $\boldsymbol{\mu} \in \mathbb{R}^{pV}$ by sampling each component $\mu_j \sim \text{Uniform}\,(-5, 5)$ for $1 \leq j \leq pV$.
2. Randomly generate a symmetric, positive-definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{pV \times pV}$.
3. Randomly generate the design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times pV}$ by sampling each row $\boldsymbol{X}_{i,:} \sim \mathcal{N}_{pV}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ for $1 \leq i \leq N$.
4. Construct view-specific feature vectors $\boldsymbol{x}_i^v = \boldsymbol{X}_{i,(1+p(v-1)):pv}$ for $1 \leq i \leq N$ and $1 \leq v \leq V$.
5. Let $h : \mathbb{R}^{pV} \rightarrow \mathbb{R}^K$ and $g : \mathbb{R}^K \rightarrow \mathbb{R}$ be randomly initialised MLPs with ReLU nonlinearities.
6. Compute $c_{i,k} = \mathbf{1}_{\left\{h\left(\boldsymbol{X}_{i,:}\right)_k \geq m_k\right\}}$, where $m_k = \text{median}\left(\left\{h\left(\boldsymbol{X}_{l,:}\right)_k\right\}_{l=1}^{N}\right)$, for $1 \leq i \leq N$ and $1 \leq k \leq K$.
7. Compute labels as $y_i = \mathbf{1}_{\left\{g\left(\boldsymbol{c}_i\right) \geq m_y\right\}}$, where $m_y = \text{median}\left(\left\{g\left(\boldsymbol{c}_i\right)\right\}_{l=1}^{N}\right)$, for $1 \leq i \leq N$.

The procedure above produces a dataset on N triples $\left(\left\{\boldsymbol{x}_i^v\right\}_{v=1}^{V}, \boldsymbol{c}_i, y_i\right)$, where $1 \leq i \leq N$. In contrast to the general setting outlined in Section 5.1 and the pediatric appendicitis dataset (Section 5.3.2), this procedure generates the same number of views for every data point. In our experiments, we set $N = 8000$, $p = 500$, $V = 3$, and $K = 30$, holding out 2000 data points as the test set.

C.2 MULTIVIEW ANIMALS WITH ATTRIBUTES

As explained in Section 5.3.1, we adapt the Animals with Attributes 2 dataset, a natural-image benchmark for attribute-based classification, to the multiview learning scenario. In particular, for each original image, we construct multiple "views" by cropping 60×60 pixel patches. As a result, concepts are only partially observable from individual views (Section 5.1). However, note that the views are not ordered, and their number is constant across all data points. Figure C.1 shows a few examples of multiview images from the MVAwA dataset.

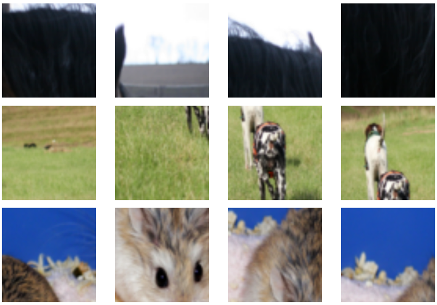


FIGURE C.1: Three examples of the four-view data points from the multiview AwA dataset. Each row corresponds to a single data point. Every view (column) constitutes a randomly chosen 60×60 px patch of the original AwA image. Observe that some concepts can be identified only from certain views, e. g., in the bottom row, attributes referring to the background cannot be detected from the second (counting from the left) view.

C.3 PEDIATRIC APPENDICITIS DATASET

The pediatric appendicitis dataset (Section 5.3.2) [332] was acquired in the course of the study approved by the Ethics Committee of the University of Regensburg (no. 18-1063-101, 18-1063_1-101, and 18-1063_2-101) following applicable guidelines and regulations. The ethics committee confirmed that there was no need for written informed consent for the retrospective analysis and publication of anonymised routine data according to Art. 27 para. 4 of the Bavarian Hospital Law. For patients followed up after discharge, written informed consent was obtained from parents or legal representatives.

	Name	Description	Pos., %
c_1	Visibility of the appendix	visibility of the vermiform appendix during the examination	76
c_2	Free intraperitoneal fluid	free fluids in the abdomen	43
c_3	Appendix layer structure	characterisation of the appendix layers, e. g. irregular in case of an increasing inflammation	14
c_4	Target sign	axial image of the appendix with the fluid-filled centre surrounded by echogenic mucosa and submucosa and hypoechoic muscularis	13
c_5	Surrounding tissue reaction	inflammation signs in tissue surrounding the appendix	33
c_6	Pathological lymph nodes	enlarged and inflamed intra-abdominal lymph nodes	21
c_7	Thickening of the bowel wall	edema of the intestinal wall, > 2–3 mm	8
c_8	Coprostasis	fecal impaction in the colon	6
c_9	Meteorism	accumulation of gas in the intestine	15

Table C.1: Explanation and descriptive statistics for the concept variables from the pediatric appendicitis dataset. Note that all variables are binary. The right-most column reports the percentage of the positive findings.

This dataset includes US images accompanied by high-level attributes. We identified relevant variables and utilised them as concepts based on two criteria: (i) the attribute has to be detectable from ultrasound images, as confirmed by a qualified physician, and (ii) the variable must be collected preoperatively. Table C.1 lists the chosen concept variables. A comprehensive summary with detailed explanations of *all* variables is available online at `http://bit.ly/3SoA5E5`.

C.4 NETWORK ARCHITECTURES

Table C.2 below provides an outline of the MVCBM architectures utilised in our experiments. Here, K denotes the number of concepts, H is the number of units in the hidden layer of g_ψ (Figure 5.1), and N_o is the number of output units dependent on the number of classes. Tables C.2a and C.2b show the architectures for the synthetic tabular and MVAwA and pediatric appendicitis datasets, respectively. Herein, `Dropout(rate)` denotes a dropout layer [416] with a probability of `rate` to drop a unit, and `BatchNorm1D()` is batch normalisation.

Notably, the architecture of $\boldsymbol{h}_{\omega^c}$ differs: we use a fully connected network for the synthetic data and the ResNet-18 backbone [339] for images. For the synthetic and MVAwA datasets, we fix $H = 100$, and for the appendicitis data, we set it to 5. For MVAwA, the output layer has $N_o = 50$ units with `softmax` activation. Since all labels in the pediatric appendicitis dataset are binary, we set $N_o = 1$ and utilise `sigmoid` activation. Lastly, for the SSMVCBM, we utilise architectures similar to those from Table C.2 for both concept prediction and representation learning "branches" (Figure 5.1).

(a)

Module	Layers
h_{ω^c}	`Linear(256); Dropout(0.05); BatchNorm1D()`
	`Linear(256); Dropout(0.05); BatchNorm1D()`
	`Linear(256); Dropout(0.05); BatchNorm1D()`
	`Linear(128)`
r_{ξ^c}	`LSTM()/mean()`
s_{χ^c}	`Linear(256); ReLU()`
	`Linear(64); ReLU()`
	`Linear(K); sigmoid()`
g_{ψ}	`Linear(H); ReLU()`
	`Linear(1); sigmoid()`

(b)

Module	Layers
h_{ω^c}	`ResNet-18()`
r_{ξ^c}	`LSTM()/mean()`
s_{χ^c}	`Linear(256); ReLU()`
	`Linear(64); ReLU()`
	`Linear(K); sigmoid()`
g_{ψ}	`Linear(H); ReLU()`
	`Linear(N_o); sigmoid()/softmax()`

TABLE C.2: Summary of the MVCBM module architectures for the (a) synthetic and (b) MVAwA and pediatric appendicitis datasets. Herein, K is the number of concepts, H denotes the number of units in the hidden layer of g_{ψ}, and N_o is the number of output units.

C.5 FURTHER RESULTS

This appendix includes supplementary results for Section 5.4. In particular, we report concept prediction AUROCs and AUPRs for the models trained with the management and severity as the target in Tables C.3 and C.4, respectively.

(a)

Model	Concept AUROC								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
CBM-seq	0.51±0.05	0.54±0.07	0.63±0.05*	0.49±0.07	0.65±0.07*	0.56±0.06	0.47±0.10	0.60±0.10	0.54±0.07
CBM-joint	0.54±0.08	0.51±0.08	0.64±0.06*	0.49±0.06	*0.67±0.03**	0.54±0.07	0.49±0.07	0.56±0.10	0.47±0.09
MVCBM-seq-avg	0.62±0.06*	0.48±0.07	0.69±0.03*	0.54±0.12	0.49±0.08	0.60±0.07*	0.48±0.09	0.47±0.13	0.57±0.09
MVCBM-seq-LSTM	**0.86±0.05***	*0.55±0.05*	0.62±0.05*	*0.69±0.03**	0.66±0.04*	**0.65±0.06***	*0.50±0.07*	**0.75±0.09***	**0.74±0.06***
MVCBM-joint-avg	0.52±0.07	0.53±0.06	*0.71±0.07**	0.59±0.05*	0.64±0.07*	**0.65±0.04***	0.48±0.10	0.54±0.07	0.52±0.15
MVCBM-joint-LSTM	0.80±0.05*	0.41±0.08	0.66±0.07*	0.61±0.04*	0.66±0.03*	*0.62±0.07**	**0.51±0.07**	0.62±0.11	0.63±0.08*
SSMVCBM-avg	0.62±0.07*	**0.57±0.08**	**0.73±0.04***	0.63±0.05*	0.55±0.04	**0.65±0.07***	*0.50±0.08*	0.49±0.08	0.52±0.05
SSMVCBM-LSTM	*0.84±0.02**	0.54±0.05	0.70±0.05*	**0.70±0.03***	**0.68±0.05***	*0.62±0.07**	*0.50±0.10*	*0.72±0.05**	*0.72±0.10**

(b)

Model	Concept AUPR								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Random	0.72	0.49	0.19	0.23	0.51	0.26	0.16	0.13	0.14
CBM-seq	0.76±0.03	*0.55±0.07*	0.37±0.09*	0.23±0.03	*0.66±0.07**	0.35±0.10	*0.19±0.06*	0.20±0.13	0.17±0.03
CBM-joint	0.77±0.04*	0.51±0.06	**0.45±0.08***	0.24±0.07	0.64±0.04*	0.29±0.04	*0.19±0.05*	0.17±0.09	0.15±0.06
MVCBM-seq-avg	0.79±0.04*	0.52±0.08	0.35±0.04*	0.31±0.14	0.51±0.06	0.37±0.08*	0.17±0.04	0.12±0.04	0.18±0.05
MVCBM-seq-LSTM	**0.95±0.02***	*0.55±0.03**	0.32±0.08*	**0.38±0.04***	*0.66±0.03**	*0.38±0.09**	0.16±0.02	**0.30±0.16**	**0.30±0.06***
MVCBM-joint-avg	0.71±0.04	0.53±0.05	0.36±0.10*	0.28±0.03*	0.60±0.07*	**0.39±0.06***	0.17±0.05	0.20±0.07	0.21±0.10
MVCBM-joint-LSTM	0.91±0.03*	0.44±0.05	0.31±0.06*	0.33±0.06*	0.64±0.03*	*0.38±0.06**	*0.19±0.04*	0.19±0.11	*0.28±0.14*
SSMVCBM-avg	0.78±0.06	**0.60±0.07***	*0.41±0.08**	0.33±0.08*	0.55±0.05	**0.39±0.07***	**0.22±0.06**	0.12±0.02	0.23±0.08
SSMVCBM-LSTM	*0.93±0.01**	*0.55±0.06*	0.38±0.09*	*0.37±0.06**	**0.67±0.06***	0.35±0.06*	0.17±0.05	*0.24±0.05**	0.27±0.08*

Table C.3: Concept prediction performance on the pediatric appendicitis dataset with the *management* as the target variable. (a) AUROCs and (b) AUPRs are reported as averages and standard deviations across ten independent initialisations.

(a)

Model	**Concept AUROC**								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
CBM-seq	0.51±0.04	*0.58±0.06**	0.61±0.08*	0.52±0.09	0.62±0.04*	0.62±0.05*	0.47±0.09	0.57±0.11	0.50±0.08
CBM-joint	0.55±0.06	0.46±0.06	0.66±0.06*	0.47±0.06	*0.64±0.04**	0.53±0.07	0.50±0.07	0.58±0.10*	0.49±0.04
MVCBM-seq-avg	0.54±0.08	0.55±0.04	**0.72±0.07***	0.62±0.04*	0.50±0.05	0.64±0.06*	0.51±0.10	0.47±0.11	0.54±0.10
MVCBM-seq-LSTM	**0.82±0.04***	0.53±0.04	0.62±0.04*	**0.69±0.04***	0.62±0.05*	**0.72±0.05***	**0.64±0.06***	**0.78±0.03***	**0.70±0.06***
MVCBM-joint-avg	0.54±0.09	0.51±0.06	0.70±0.06*	0.59±0.08*	0.61±0.06*	0.62±0.05*	0.54±0.15	0.48±0.14	0.55±0.12
MVCBM-joint-LSTM	**0.82±0.03***	0.48±0.06	0.66±0.07*	0.64±0.06*	**0.65±0.05***	0.64±0.09*	0.47±0.09	0.61±0.14	*0.65±0.05**
SSMVCBM-avg	0.53±0.06*	0.56±0.08*	*0.71±0.05**	0.60±0.06*	0.51±0.05	0.64±0.09*	0.46±0.08	0.48±0.09	0.53±0.03
SSMVCBM-LSTM	*0.77±0.10**	**0.59±0.08**	0.70±0.06*	*0.67±0.07**	**0.65±0.07***	*0.67±0.05**	*0.62±0.08**	*0.74±0.15**	0.64±0.11*

(b)

Model	**Concept AUPR**								
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Random	0.72	0.49	0.19	0.23	0.51	0.26	0.16	0.13	0.14
CBM-seq	0.75±0.03	*0.58±0.05**	0.34±0.09*	0.24±0.05	0.64±0.04*	0.35±0.06*	0.18±0.05	0.19±0.07	0.15±0.03
CBM-joint	0.77±0.05	0.47±0.04	*0.37±0.09**	0.25±0.06	0.64±0.05*	0.30±0.07	0.17±0.04	0.18±0.06	0.18±0.08
MVCBM-seq-avg	0.75±0.05	*0.58±0.06**	**0.42±0.07***	0.33±0.06*	0.53±0.05	0.41±0.08*	0.21±0.05	0.13±0.05	*0.24±0.12*
MVCBM-seq-LSTM	*0.91±0.04**	0.55±0.04*	0.33±0.08*	**0.40±0.06***	*0.65±0.03**	**0.50±0.11***	*0.23±0.05**	*0.27±0.05**	**0.26±0.07***
MVCBM-joint-avg	0.74±0.06	0.51±0.07	**0.42±0.09***	0.28±0.07	0.59±0.06*	0.35±0.05*	0.22±0.06	0.22±0.13	0.21±0.08
MVCBM-joint-LSTM	**0.92±0.02***	0.49±0.05	*0.37±0.11**	0.32±0.07*	*0.65±0.06**	0.39±0.07*	0.20±0.06	0.17±0.07	0.21±0.06*
SSMVCBM-avg	0.73±0.05	*0.58±0.07**	0.36±0.05*	0.28±0.04*	0.53±0.05	0.37±0.09*	0.20±0.06	0.13±0.02	*0.24±0.06**
SSMVCBM-LSTM	0.88±0.06*	**0.60±0.06***	**0.42±0.06***	*0.39±0.09**	**0.67±0.07***	*0.43±0.10**	**0.24±0.08**	**0.30±0.13***	0.20±0.05*

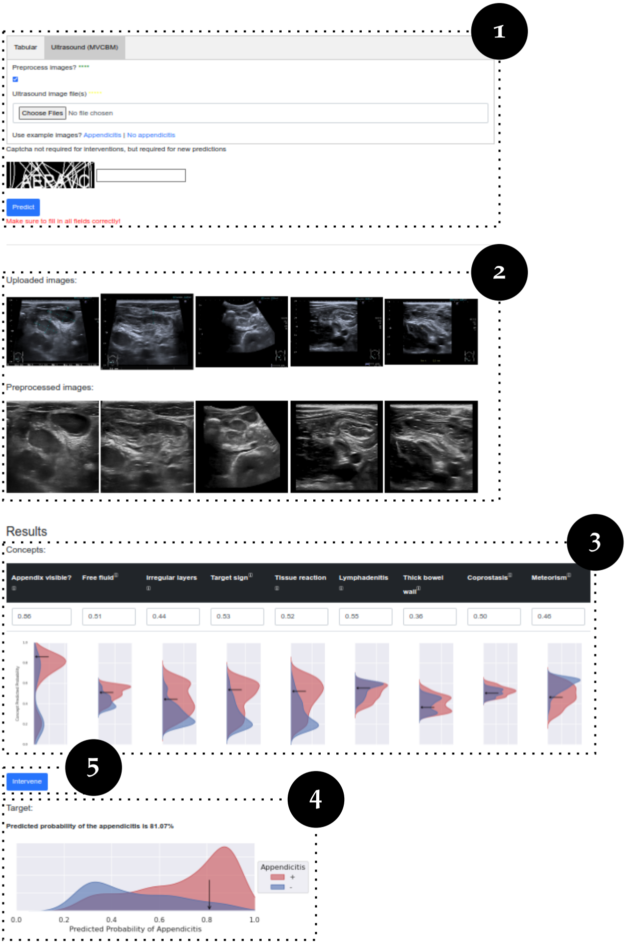
TABLE C.4: Concept prediction performance on the pediatric appendicitis dataset with the *severity* as the target variable. (a) AUROCs and (b) AUPRs are reported as averages and standard deviations across ten independent initialisations.

C.6 ONLINE PREDICTION TOOL

Below, we provide further details on our online pediatric appendicitis prediction tool. In its current version, it is a research prototype and should be utilised solely for noncommercial, educational purposes and *not* for clinical decision-making.

The tool deploys a multiview concept bottleneck model trained to predict the diagnosis using the sequential optimisation procedure and LSTM to fuse the views (MVCBM-seq-LSTM from Table 5.3). We use parameters learnt after training from one of the initialisations included in our experiments. Note that the model *was not* retrained on the complete dataset.

Figure C.2 provides a concrete example of how the tool may be used. The user uploads ultrasound images. If requested, UI element regions are masked and filled, and CLAHE is applied. Then, processed images are forwarded to the trained MVCBM network, which predicts the concept values and the diagnosis label and displays them. The user may intervene if they choose and recalculate the final prediction using adjusted concept values.



Preprocess images? ****

☑

Ultrasound image file(s) *****

No file chosen

Use example images? Appendicitis | No appendicitis

Captcha not required for interventions, but required for new predictions

Make sure to fill in all fields correctly!

Uploaded images:

Preprocessed images:

Results

Concepts:

Appendix visible?	Free fluid	Irregular layers	Target sign	Tissue reaction	Lymphadenitis	Thick bowel wall	Coprostasis	Meteorism
0.56	0.51	0.44	0.53	0.52	0.55	0.36	0.50	0.46

Target:

Predicted probability of the appendicitis is 81.07%

FIGURE C.2: A summary of the pediatric appendicitis online prediction tool. ❶ The user uploads ultrasound images for a single patient. ❷ Optionally, image preprocessing is performed. ❸ The tool displays predicted concept values given by sigmoid activations alongside predicted value histograms obtained from the *training* data (plotted separately for appendicitis and non-appendicitis cases in red and blue, respectively). ❹ The tool shows the prediction for the diagnosis. ❺ The user may intervene by editing concept predictions and, thus, affect the target prediction. Upon intervention, an updated predicted value is displayed.

D

SUPPLEMENTARY: BEYOND CONCEPT BOTTLENECKS

D.1 INTERVENTION STRATEGIES

In one of the ablation experiments in Chapter 6 (Section 6.3.2), we compare intervention strategies inspired by Shin *et al.* [348]: (i) random-subset and (ii) uncertainty-based. Algorithms D.1.1–D.1.2 describe these procedures.

Algorithm D.1.1: Random-subset Intervention Strategy

```
Input: A data point $(\boldsymbol{x}, \boldsymbol{c}, y)$; predicted concept values $\hat{\boldsymbol{c}}$; number of
       concept variables $1 \leq k \leq K$ to be intervened on
Output: Intervened concept values $\boldsymbol{c}'$

Let $\boldsymbol{c}' \leftarrow \hat{\boldsymbol{c}}$
Sample $\mathcal{I}$ uniformly at random from $\{\mathcal{S} \subseteq \{1, \dots, K\} : |\mathcal{S}| = k\}$
Assign $\boldsymbol{c}'_{\mathcal{I}} \leftarrow \boldsymbol{c}_{\mathcal{I}}$

return $\boldsymbol{c}'$
```

Algorithm D.1.2: Uncertainty-based Intervention Strategy

```
Input: A data point $(\boldsymbol{x}, \boldsymbol{c}, y)$; predicted concept values $\hat{\boldsymbol{c}}$; number of
       concept variables $1 \leq k \leq K$ to be intervened on
Output: Intervened concept values $\boldsymbol{c}'$

Compute $\sigma_j \leftarrow 1 / \left(\left|\hat{c}_j - 0.5\right| + \varepsilon\right)$ for $1 \leq j \leq K$, where $\varepsilon > 0$ is small
Let $\boldsymbol{\sigma} \leftarrow \begin{pmatrix} \sigma_1 & \cdots & \sigma_K \end{pmatrix}$
Let $\boldsymbol{c}' \leftarrow \hat{\boldsymbol{c}}$
Sample $k$ indices $\mathcal{I} = \{i_j\}_{j=1}^{k}$ s.t. each $i_j$ is sampled without
  replacement from $\{1, \dots, K\}$ with initial probabilities given by
  $(\boldsymbol{\sigma} + \varepsilon) \,/\, \left(K\varepsilon + \sum_{i=1}^{K} \sigma_i\right)$, where $\varepsilon > 0$ is small
$\boldsymbol{c}'_{\mathcal{I}} \leftarrow \boldsymbol{c}_{\mathcal{I}}$

return $\boldsymbol{c}'$
```

Recall that given a data point $(\boldsymbol{x}, \boldsymbol{c}, y)$ and predicted values $\hat{\boldsymbol{c}}$ and $\hat{y}$, an intervention strategy defines a distribution π over intervened concept values $\boldsymbol{c}'$ (Section 6.2.1). The **random-subset strategy** (Algorithm D.1.1) replaces predicted values with the ground truth for $1 \leq k \leq K$ concept variables chosen uniformly at random. By contrast, the **uncertainty-based** strategy (Algorithm D.1.2) samples concept variables to be replaced without replacement with initial probabilities proportional to the concept prediction uncertainties, denoted by $\boldsymbol{\sigma}$. In our experiments, the components of $\hat{\boldsymbol{c}}$ are the outputs of the sigmoid function, and the uncertainties are computed as $\sigma_i = 1/\left(|\hat{c}_i - 0.5| + \varepsilon\right)$ [348] for $1 \leq i \leq K$, where $\varepsilon > 0$ is small.

D.2 SYNTHETIC DATA GENERATION

As explained in Section 6.3.1, throughout Chapter 6, we consider three generating mechanisms for synthetic nonlinear tabular data: (i) *bottleneck*, (ii) *confounder*, and (iii) *incomplete*. In this appendix, we comprehensively describe the three procedures, following the notation introduced in Sections 5.1 and 6.1.

In the ***bottleneck*** scenario (Figure 6.2a), the covariates $\boldsymbol{x}_i$ generate binary concepts $\boldsymbol{c}_i \in \{0, 1\}^K$, and the binary target y_i depends on the covariates exclusively via the concepts. The generative process is as follows:

1. Randomly sample $\boldsymbol{\mu} \in \mathbb{R}^p$ s.t. $\mu_j \sim \text{Uniform}\left(-5, 5\right)$ for $1 \leq j \leq p$.
2. Generate a random symmetric, positive-definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$.
3. Randomly sample a design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ s.t. $\boldsymbol{X}_{i,:} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$.
4. Let $h : \mathbb{R}^p \rightarrow \mathbb{R}^K$ and $g : \mathbb{R}^K \rightarrow \mathbb{R}$ be randomly initialised multilayer perceptrons with ReLU nonlinearities.
5. Let $c_{i,k} = \mathbf{1}_{\left\{h\left(\boldsymbol{X}_{i,:}\right)_k \geq m_k\right\}}$, where $m_k = \text{median}\left(\left\{h\left(\boldsymbol{X}_{l,:}\right)_k\right\}_{l=1}^N\right)$, for $1 \leq i \leq N$ and $1 \leq k \leq K$.
6. Let $y_i = \mathbf{1}_{\left\{g\left(\boldsymbol{c}_i\right) \geq m_y\right\}}$, where $m_y = \text{median}\left(\left\{g\left(\boldsymbol{c}_l\right)\right\}_{l=1}^N\right)$, for $1 \leq i \leq N$.

In another setting we consider, $\boldsymbol{x}$ and $\boldsymbol{c}$ are generated by an unobserved ***confounder*** (Figure 6.2b):

1. Randomly sample $\boldsymbol{U} \in \mathbb{R}^{N \times K}$ s.t. $u_{i,k} \sim \mathcal{N}(0, 1)$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.
2. Let $c_{i,k} = \mathbf{1}_{\{u_{i,k} \geq 0\}}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.
3. Let $h : \mathbb{R}^K \rightarrow \mathbb{R}^p$ and $g : \mathbb{R}^K \rightarrow \mathbb{R}$ be randomly initialised multilayer perceptrons with ReLU nonlinearities.
4. Let $\boldsymbol{x}_i = h\left(\boldsymbol{U}_{i,:}\right)$ for $1 \leq i \leq N$.
5. Let $y_i = \mathbf{1}_{\{\mathrm{sigmoid}(g(\boldsymbol{c}_i)) \geq 1/2\}}$ for $1 \leq i \leq N$.

Lastly, to investigate the ***incomplete*** concept set scenario (Figure 5.2, Section 5.2.3), we slightly adjust the procedure from the *bottleneck* setting above by making a subset of concepts latent:

1. Follow steps 1–3 from the *bottleneck* procedure.
2. Let $h : \mathbb{R}^p \rightarrow \mathbb{R}^{K+J}$ and $g : \mathbb{R}^{K+J} \rightarrow \mathbb{R}$ be randomly initialised multilayer perceptrons with ReLU nonlinearities, where J is the number of unobserved concept variables.
3. Let $u_{i,k} = \mathbf{1}_{\left\{h\left(\boldsymbol{X}_{i,:}\right)_k \geq m_k\right\}}$, where $m_k = \mathrm{median}\left(\left\{h\left(\boldsymbol{X}_{l,:}\right)_k\right\}_{l=1}^N\right)$, for $1 \leq i \leq N$ and $1 \leq k \leq K + J$.
4. Let $\boldsymbol{c}_i = \boldsymbol{u}_{i,1:K}$ and $\boldsymbol{r}_i = \boldsymbol{u}_{i,(K+1):(K+J)}$ for $1 \leq i \leq N$.
5. Let $y_i = \mathbf{1}_{\left\{g(\boldsymbol{u}_i) \geq m_y\right\}}$, where $m_y = \mathrm{median}\left(\left\{g\left(\boldsymbol{u}_i\right)\right\}_{l=1}^N\right)$, for $1 \leq i \leq N$.

Observe that above $\boldsymbol{u}_i$ corresponds to the concatenation of $\boldsymbol{c}_i$ and $\boldsymbol{r}_i$. Across all experiments in Chapter 6, we set $J = 90$.

D.3 NETWORK ARCHITECTURES

In our experiments on synthetic tabular data (Section 6.4.1), we utilise an MLP as the black-box classifier. Its architecture is summarised in Table D.1. For this classifier, probing functions are trained and interventions are performed on the third layer, i. e. the output after line **2** in Table D.1.

For natural and medical image datasets, we use the ResNet-18 [339] with random initialisation followed by four fully connected layers and the sigmoid or softmax activation. Probing and interventions are performed on the activations of the second layer after the ResNet-18 backbone.

For CBMs, to facilitate fair comparison, we use the same architectures with the exception that the layers mentioned above are converted into bottlenecks with appropriate dimensionality and activation functions. Similar settings are used for *post hoc* CBMs with the addition of a linear layer mapping representations to the concepts.

Lastly, during fine-tuning, we utilise a single fully connected layer with an appropriate activation function as a *linear* probe and a multilayer perceptron with a single hidden layer as a *nonlinear* probing function.

Black box

```
1   Linear(256); ReLU()
    Dropout(0.05)
    BatchNorm1D()
2   for l in range(2):
        Linear(256); ReLU()
        Dropout(0.05)
        BatchNorm1D()
3   Linear(1); sigmoid()
```

Table D.1: Neural network architecture of the black-box classifier from the experiments on the synthetic tabular data.

E

SUPPLEMENTARY: INTERPLAY BETWEEN EXPLANATION AND FAIRNESS

E.1 DECISION BOUNDARY COVARIANCE

Below, we examine the relationship between the classification parity proxy functions introduced in Section 7.2.1 and the (conditional) covariance between the decision boundary of the given classifier $f_{\boldsymbol{\theta}}$ and the protected attribute a.

Following the setting in Section 7.2.1, let us assume being given a dataset $\mathscr{D} = \{(\boldsymbol{x}_i, a_i, y_i)\}_{i=1}^{N}$. Furthermore, let $K = \sum_{i=1}^{N} a_i$ be the number of data points with $a_i = 1$ and let $\overline{f_{\boldsymbol{\theta}}(\boldsymbol{x})} = \frac{1}{N}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ denote the average output of the classifier. Recall that the sample covariance between $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and a is given by

$$\widehat{\mathrm{Cov}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}), a\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \overline{f_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)\left(a_i - \frac{K}{N}\right) \tag{E.1}$$

$$= \frac{1}{N-1}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, a_i - \frac{K}{N-1}\overline{f_{\boldsymbol{\theta}}(\boldsymbol{x})}. \tag{E.2}$$

Lemma E.1.1. The proxy function for the SPD from Equation 7.3 is proportional to the empirical estimator of the covariance between $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and a: $-\tilde{\mu}_{\mathrm{SPD}}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right) \propto \widehat{\mathrm{Cov}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}), a\right)$.

Proof. Starting from Equation 7.3, observe that

$$-\tilde{\mu}_{\mathrm{SPD}}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right) = \frac{\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, a_i}{\sum_{i=1}^{N} a_i} - \frac{\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)(1-a_i)}{\sum_{i=1}^{N} 1 - a_i} \tag{E.3}$$

$$= \frac{1}{K}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, a_i - \frac{N}{N-K}\overline{f_{\boldsymbol{\theta}}(\boldsymbol{x})} + \frac{1}{N-K}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, a_i \tag{E.4}$$

$$= \frac{N}{K(N-K)}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\, a_i - \frac{NK}{K(N-K)}\overline{f_{\boldsymbol{\theta}}(\boldsymbol{x})}. \tag{E.5}$$

Note that the expression in Equation E.5 is proportional to Equation E.2 by a factor of $\frac{K(N-K)}{N(N-1)}$, constant in the model's parameters $\boldsymbol{\theta}$. ☐

A similar analysis can be performed for the proxy function of the EOD (Equation 7.4). By contrast, in this case, we examine the *conditional* covariance between $f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)$ and a:

$$\operatorname{Cov}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right), a \middle| y=1\right) = \mathbb{E}\Big[\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \mathbb{E}\left[f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)|y=1\right]\right) \tag{E.6}$$

$$\cdot\left(a - \mathbb{E}\left[a|y=1\right]\right)\Big|y=1\Big] \tag{E.7}$$

$$= \mathbb{E}\left[f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)a|y=1\right] - \mathbb{E}\left[f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)|y=1\right]\mathbb{E}\left[a|y=1\right], \tag{E.8}$$

where Equation E.8 can be derived from Equations E.6–E.7 using the properties of conditional expectations. In addition to the notation introduced above, let $M = \sum_{i=1}^{N} y_i$ be the number of positive cases and $R = \sum_{i=1}^{N} a_i y_i$ be the number of positive cases with $a_i = 1$. We will consider the following empirical estimate of the conditional covariance above, which is obtained by plugging in consistent estimators for conditional expectations:

$$\widehat{\operatorname{Cov}}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right), a|y=1\right) = \frac{\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)a_i y_i}{\sum_{i=1}^{N} y_i} - \frac{\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)y_i}{\sum_{i=1}^{N} y_i}\cdot\frac{\sum_{i=1}^{N} a_i y_i}{\sum_{i=1}^{N} y_i} \tag{E.9}$$

$$= \frac{1}{M}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)a_i y_i - \frac{R}{M^2}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)y_i. \tag{E.10}$$

Lemma E.1.2. The proxy function for the EOD from Equation 7.4 is proportional to the empirical estimator in Equation E.10:
$-\tilde{\mu}_{\mathrm{EOD}}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right) \propto \widehat{\operatorname{Cov}}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right), a|y=1\right)$.

Proof. Using Equation 7.4, observe that

$$-\tilde{\mu}_{\mathrm{EOD}}\left(f_{\boldsymbol{\theta}}, \mathscr{D}\right) = \frac{\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)a_i y_i}{\sum_{i=1}^{N} a_i y_i} - \frac{\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\left(1-a_i\right)y_i}{\sum_{i=1}^{N}\left(1-a_i\right)y_i} \tag{E.11}$$

$$= \frac{1}{R}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)a_i y_i - \frac{1}{M-R}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)y_i + \frac{1}{M-R}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)a_i y_i \tag{E.12}$$

$$= \frac{M}{R\left(M-R\right)}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)a_i y_i - \frac{1}{M-R}\sum_{i=1}^{N} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)y_i. \tag{E.13}$$

The expression in Equation E.13 is proportional to Equation E.10 by a factor of $\frac{R(M-R)}{M^2}$, constant in the model's parameters $\boldsymbol{\theta}$. □

E.2 NETWORK ARCHITECTURES

Table E.1 contains a pseudocode description of the fully connected neural network architecture for the classifier in our debiasing experiments on tabular data (Section 7.3.2).

Classifier

```
1   Linear(32); ReLU()
    Dropout(0.05)
    BatchNorm1D()
2   for l in range(10):
        Linear(32); ReLU()
        Dropout(0.05)
        BatchNorm1D()
3   Linear(1); sigmoid()
```

Table E.1: Fully connected neural network architecture used as the classifier in debiasing experiments on tabular data.

BIBLIOGRAPHY

[1] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams, "Dynabench: Rethinking benchmarking in NLP", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds., Association for Computational Linguistics, 2021, 4110. DOI: `10.18653/v1/2021.naacl-main.324`.

[2] Y. LeCun and C. Cortes, *MNIST handwritten digit database*, `http://yann.lecun.com/exdb/mnist/`, 2010.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database", in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 248. DOI: `10.1109/CVPR.2009.5206848`.

[4] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, 5, 2001. DOI: `10.1023/a:1010933404324`.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *The Annals of Statistics*, vol. 29, no. 5, 1189, 2001. DOI: `10.1214/aos/1013203451`.

[6] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Networks*, vol. 61, 85, 2015. DOI: `10.1016/j.neunet.2014.09.003`.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[8] R. Bommasani *et al.*, *On the opportunities and risks of foundation models*, arXiv:2108.07258, 2021. DOI: `10.48550/arXiv.2108.07258`.

[9] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence", *Nature*, vol. 616, no. 7956, 259, 2023. DOI: `10.1038/s41586-023-05881-4`.

[10] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)", *Statistical Science*, vol. 16, no. 3, 199, 2001. DOI: `10.1214/ss/1009213726`.

[11] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, vol. 1, no. 5, 206, 2019. DOI: `10.1038/s42256-019-0048-x`.

[12] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR", *Harvard Journal of Law & Technology*, vol. 31, 2 2017. DOI: `10.2139/ssrn.3063289`.

[13] Z. C. Lipton, "The mythos of model interpretability", *Queue*, vol. 16, no. 3, 31, 2018. DOI: `10.1145/3236386.3241340`.

[14] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang, *Causality learning: A new perspective for interpretable machine learning*, arXiv:2006.16789, 2020. DOI: `10.48550/arXiv.2006.16789`.

[15] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, arXiv:1702.08608, 2017. DOI: `10.48550/arXiv.1702.08608`.

[16] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions", *Nature Communications*, vol. 10, no. 1, 2019. DOI: `10.1038/s41467-019-12875-2`.

[17] K. Liu, N. Sadoune, N. Rao, J. Greitemann, and L. Pollet, "Revealing the phase diagram of Kitaev materials by machine learning: Cooperation and competition between spin liquids", *Physical Review Research*, vol. 3, 023016, 2 2021. DOI: `10.1103/PhysRevResearch.3.023016`.

[18] T. Pimentel, A. D. McCarthy, D. Blasi, B. Roark, and R. Cotterell, "Meaning to form: Measuring systematicity as information", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, 2019, 1751. DOI: `10.18653/v1/P19-1171`.

[19] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems", *Machine Learning*, vol. 102, no. 3, 349, 2015. DOI: `10.1007/s10994-015-5528-6`.

[20] ——, "Optimized risk scores", in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, 2017, 1125. DOI: `10.1145/3097983.3098161`.

[21] F. Wang and C. Rudin, "Falling Rule Lists", in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, G. Lebanon and S. V. N. Vishwanathan, Eds., ser. Proceedings of Machine Learning Research, vol. 38, San Diego, California, USA: PMLR, 2015, 1013.

[22] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models", in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, 5338.

[23] M. Lee, M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, X. L. Li, F. Ladhak, F. Rong, R. E. Wang, M. Kwon, J. S. Park, H. Cao, T. Lee, R. Bommasani, M. S. Bernstein, and P. Liang, "Evaluating human-language model interaction", *Transactions on Machine Learning Research*, 2023.

[24] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig, "Explain, edit, and understand: Rethinking user study design for evaluating model explanations", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 5, 5277, 2022. DOI: `10.1609/aaai.v36i5.20464`.

[25] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics", *Electronics*, vol. 8, no. 8, 832, 2019. DOI: `10.3390/electronics8080832`.

[26] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission", in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15, Sydney, NSW, Australia: Association for Computing Machinery, 2015, 1721. DOI: `10.1145/2783258.2788613`.

[27] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, *How we analyzed the COMPAS recidivism algorithm*, `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`, 2016.

[28] A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, "xxAI - beyond explainable artificial intelligence", in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham: Springer International Publishing, 2022, 3. DOI: `10.1007/978-3-031-04083-2_1`.

[29] R. B. Altman and M. Levitt, "What is biomedical data science and do we need an annual review of it?", *Annual Review of Biomedical Data Science*, vol. 1, i, 2018. DOI: `10.1146/annurev-bd-01-041718-100001`.

[30] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 2005. DOI: `10.1007/978-3-540-27752-1`.

[31] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future", *npj Digital Medicine*, vol. 5, no. 1, 2022. DOI: `10.1038/s41746-022-00592-y`.

[32] C. Xu, D. Tao, and C. Xu, *A survey on multi-view learning*, arXiv:1304.5634, 2013. DOI: `10.48550/arXiv.1304.5634`.

[33] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, 423, 2019. DOI: `10.1109/TPAMI.2018.2798607`.

[34] F. Imrie, R. Davis, and M. van der Schaar, "Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare", *Nature Machine Intelligence*, vol. 5, no. 8, 824, 2023. DOI: `10.1038/s42256-023-00698-2`.

[35] T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning", in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, 2009, 9. DOI: `10.1007/978-0-387-84858-7_2`.

[36] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, *A cookbook of self-supervised learning*, arXiv:2304.12210, 2023. DOI: `10.48550/arXiv.2304.12210`.

[37] Z.-H. Zhou, "A brief introduction to weakly supervised learning", *National Science Review*, vol. 5, no. 1, 44, 2017. DOI: `10.1093/nsr/nwx106`.

[38] S. Otálora, N. Marini, H. Müller, and M. Atzori, "Combining weakly and strongly supervised learning improves strong supervision in Gleason pattern classification", *BMC Medical Imaging*, vol. 21, no. 1, 2021. DOI: `10.1186/s12880-021-00609-0`.

[39] R. Marcinkevičs and J. E. Vogt, *Interpretability and explainability: A machine learning zoo mini-tour*, arXiv:2012.01805, 2020. DOI: `10.48550/arXiv.2012.01805`.

[40] ——, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples", *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 3, e1493, 2023. DOI: `10.1002/widm.1493`.

[41] C. Molnar, *Interpretable Machine Learning*. 2018, `https://christophm.github.io/interpretable-ml-book/`.

[42] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Publishing Company, Incorporated, 2019.

[43] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression", in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12, Beijing, China: Association for Computing Machinery, 2012, 150. DOI: `10.1145/2339530.2339556`.

[44] H. Allahyari and N. Lavesson, "User-oriented assessment of classification model understandability", in *11th Scandinavian Conference on Artificial Intelligence*, A. Kofod-Petersen, Ed., ser. Frontiers in Artificial Intelligence and Applications, vol. 227, Trondheim, Norway: IOS Press, 2011, 11.

[45] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions", in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA: Association for Computing Machinery, 2013, 623. DOI: `10.1145/2487575.2487579`.

[46] T. Hastie, R. Tibshirani, and J. Friedman, "Linear methods for regression", in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, 2009, 43. DOI: `10.1007/978-0-387-84858-7_3`.

[47] J. O. Ramsay, "Monotone regression splines in action", *Statistical Science*, vol. 3, no. 4, 425, 1988.

[48] T. Hastie and R. Tibshirani, "Generalized additive models", *Statistical Science*, vol. 1, no. 3, 297, 1986.

[49] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Contrastive explanations and consequential recommendations", *ACM Computing Surveys*, vol. 55, no. 5, 2022. DOI: `10.1145/3527848`.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin, "''Why should I trust you?": Explaining the predictions of any classifier", ser. KDD '16, San Francisco, California, USA: Association for Computing Machinery, 2016, 1135. DOI: `10.1145/2939672.2939778`.

[51] G. K. Dziugaite, S. Ben-David, and D. M. Roy, *Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability*, arXiv:2010.13764, 2020. DOI: `10.48550/arXiv.2010.13764`.

[52] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges", *Statistics Surveys*, vol. 16, 1, 2022. DOI: `10.1214/21-SS133`.

[53] L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models", in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, 1827. DOI: `10.1145/3531146.3533232`.

[54] T. M. Mitchell, "Learning sets of rules", in *Machine learning*, ser. McGraw-Hill Series in Computer Science. USA: McGraw-Hill, 1997, 274.

[55] L. De Raedt, "A perspective on inductive logic programming", in *The Logic Programming Paradigm: A 25-Year Perspective*, K. R. Apt, V. W. Marek, M. Truszczynski, and D. S. Warren, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, 335. DOI: `10.1007/978-3-642-60085-2_14`.

[56] W.-Y. Loh, “Classification and regression trees”, *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 1, 14, 2011. DOI: `10.1002/widm.8`.

[57] W. W. Cohen, “Fast effective rule induction”, in *Machine Learning Proceedings 1995*, A. Prieditis and S. Russell, Eds., San Francisco, CA: Morgan Kaufmann, 1995, 115. DOI: `10.1016/B978-1-55860-377-6.50023-2`.

[58] J. H. Friedman and B. E. Popescu, “Predictive learning via rule ensembles”, *The Annals of Applied Statistics*, vol. 2, no. 3, 916, 2008. DOI: `10.1214/07-AOAS148`.

[59] F. Wang and C. Rudin, *Causal falling rule lists*, arXiv:1510.05189, 2015. DOI: `10.48550/arXiv.1510.05189`.

[60] C. Chen and C. Rudin, “An optimization approach to learning falling rule lists”, in *Proceedings of the 21th International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, Eds., ser. Proceedings of Machine Learning Research, vol. 84, PMLR, 2018, 604.

[61] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, “Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today’s critically ill patients”, *Critical Care Medicine*, vol. 34, no. 5, 1297, 2006. DOI: `10.1097/01.ccm.0000215112.84523.f0`.

[62] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, and D. C. Angus, “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)”, *JAMA*, vol. 315, no. 8, 801, 2016. DOI: `10.1001/jama.2016.0287`.

[63] H. Zhang, Q. Morris, B. Ustun, and M. Ghassemi, “Learning optimal predictive checklists”, in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, 1215.

[64] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, “Optimization with sparsity-inducing penalties”, *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, 1, 2011. DOI: `10.1561/2200000015`.

[65] L. Rosacco, *Lecture notes for 9.520: Statistical learning theory and applications*, 2014.

[66] R. Tibshirani, “Regression shrinkage and selection via the lasso”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, 267, 1996. DOI: `10.1111/j.2517-6161.1996.tb02080.x`.

[67] L. Meier, S. Van De Geer, and P. Bühlmann, “The group lasso for logistic regression”, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, no. 1, 53, 2008. DOI: `10.1111/j.1467-9868.2007.00627.x`.

[68] R. Tibshirani, “The lasso method for variable selection in the Cox model”, *Statistics in Medicine*, vol. 16, no. 4, 385, 1997. DOI: `10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3`.

[69] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net”, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, 301, 2005. DOI: `10.1111/j.1467-9868.2005.00503.x`.

[70] S. Bakin, *Adaptive regression and model selection in data mining problems*, PhD Thesis, 1999. DOI: `10.25911/5D78DB4C25DBB`.

[71] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, 49, 2006. DOI: `10.1111/j.1467-9868.2005.00532.x`.

[72] J. Huang, T. Zhang, and D. Metaxas, “Learning with structured sparsity”, *Journal of Machine Learning Research*, vol. 12, no. 103, 3371, 2011.

[73] J. Frecon, S. Salzo, and M. Pontil, “Bilevel learning of the group lasso structure”, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS’18, Montréal, Canada: Curran Associates Inc., 2018, 8311.

[74] J. Guerguiev, T. P. Lillicrap, and B. A. Richards, “Towards deep learning with segregated dendrites”, *eLife*, vol. 6, P. Latham, Ed., e22901, 2017. DOI: `10.7554/eLife.22901`.

[75] A. Gelman, “Scaling regression inputs by dividing by two standard deviations”, *Statistics in Medicine*, vol. 27, no. 15, 2865, 2008. DOI: `10.1002/sim.3107`.

[76] J. A. Nelder and R. W. M. Wedderburn, “Generalized linear models”, *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, 370, 1972. DOI: `10.2307/2344614`.

[77] T. Hastie and D. Pregibon, “Generalized linear models”, in *Statistical Models in S*, J. M. Chambers and T. Hastie, Eds. Wadsworth & Brooks/Cole, 1992, 195.

[78] U. Segal, “A sufficient condition for additively separable functions”, *Journal of Mathematical Economics*, vol. 23, no. 3, 295, 1994. DOI: `10.1016/0304-4068(94)90009-4`.

[79] J. Rice and M. Rosenblatt, “Smoothing splines: Regression, derivatives and deconvolution”, *The Annals of Statistics*, vol. 11, no. 1, 141, 1983. DOI: `10.1214/aos/1176346065`.

[80] H. Liu, L. Wasserman, J. Lafferty, and P. Ravikumar, “SpAM: Sparse additive models”, in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, Curran Associates, Inc., 2007.

[81] T. Hastie and R. Tibshirani, “Varying-coefficient models”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 55, no. 4, 757, 1993. DOI: `10.1111/j.2517-6161.1993.tb01939.x`.

[82] W. S. Cleveland and E. Grosse, “Computational methods for local regression”, *Statistics and Computing*, vol. 1, no. 1, 47, 1991. DOI: `10.1007/bf01890836`.

[83] J. Fan and W. Zhang, “Statistical estimation in varying coefficient models”, *The Annals of Statistics*, vol. 27, no. 5, 1491, 1999. DOI: `10.1214/aos/1017939139`.

[84] D. Alvarez Melis and T. Jaakkola, “Towards robust interpretability with self-explaining neural networks”, in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018.

[85] M. Al-Shedivat, A. Dubey, and E. Xing, “Contextual explanation networks”, *Journal of Machine Learning Research*, vol. 21, no. 194, 1, 2020.

[86] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, “Neural additive models: Interpretable machine learning with neural nets”, in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, 4699.

[87] C.-H. Chang, R. Caruana, and A. Goldenberg, “NODE-GAM: Neural generalized additive model for interpretable deep learning”, in *International Conference on Learning Representations*, 2022.

[88] A. Sarkar, "Is explainable AI a race against model complexity?", in *Workshop on Transparency and Explanations in Smart Systems (TeXSS), in conjunction with ACM Intelligent User Interfaces (IUI 2022)*, CEUR Workshop Proceedings, 2022, 192.

[89] J. Feng and N. Simon, *Sparse-input neural networks for high-dimensional nonparametric regression and classification*, arXiv:1711.07592, 2017. DOI: 10.48550/arXiv.1711.07592.

[90] ——, "Ensembled sparse-input hierarchical networks for high-dimensional datasets", *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 6, 736, 2022. DOI: 10.1002/sam.11579.

[91] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso", *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, 231, 2013. DOI: 10.1080/10618600.2012.681250.

[92] Y. Li, C.-Y. Chen, and W. W. Wasserman, "Deep feature selection: Theory and application to identify enhancers and promoters", *Journal of Computational Biology*, vol. 23, no. 5, 322, 2016. DOI: 10.1089/cmb.2015.0189.

[93] I. Lemhadri, F. Ruan, L. Abraham, and R. Tibshirani, "LassoNet: A neural network with feature sparsity", *Journal of Machine Learning Research*, vol. 22, no. 127, 1, 2021.

[94] N. H. Choi, W. Li, and J. Zhu, "Variable selection with the strong heredity constraint and its oracle property", *Journal of the American Statistical Association*, vol. 105, no. 489, 354, 2010. DOI: 10.1198/jasa.2010.tm08281.

[95] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox, "Neural Granger causality", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, 4267, 2021. DOI: 10.1109/TPAMI.2021.3065601.

[96] S. Khanna and V. Y. F. Tan, "Economy statistical recurrent units for inferring nonlinear Granger causality", in *8th International Conference on Learning Representations, ICLR 2020*, OpenReview.net, 2020.

[97] J. Schmidhuber, "Learning to control fast-weight memories: An alternative to dynamic recurrent networks", *Neural Computation*, vol. 4, no. 1, 131, 1992. DOI: 10.1162/neco.1992.4.1.131.

[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[99] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, *Attention interpretability across NLP tasks*, arXiv:1909.11218, 2019. DOI: 10.48550/arXiv.1909.11218.

[100] P. Schwab, D. Miladinovic, and W. Karlen, "Granger-causal attentive mixtures of experts: Learning important features with neural networks", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 4846, 2019. DOI: 10.1609/aaai.v33i01.33014846.

[101] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks", *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, 312, 2019. DOI: 10.3390/make1010019.

[102] S. Jain and B. C. Wallace, "Attention is not Explanation", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, 3543. DOI: 10.18653/v1/N19-1357.

[103] S. Serrano and N. A. Smith, “Is attention interpretable?”, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, 2931. DOI: 10.18653/v1/P19-1282.

[104] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, 1798, 2013. DOI: 10.1109/TPAMI.2013.50.

[105] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes”, in *2nd International Conference on Learning Representations, ICLR 2014*, Y. Bengio and Y. LeCun, Eds., 2014.

[106] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, “β-VAE: Learning basic visual concepts with a constrained variational framework”, in *5th International Conference on Learning Representations, ICLR 2017*, OpenReview.net, 2017.

[107] F. Locatello, B. Poole, G. Raetsch, B. Schölkopf, O. Bachem, and M. Tschannen, “Weakly-supervised disentanglement without compromises”, in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, 6348.

[108] A. Taeb, N. Ruggeri, C. Schnuck, and F. Yang, *Provable concept learning for interpretable predictions using variational autoencoders*, arXiv:2204.00492, 2022. DOI: 10.48550/arXiv.2204.00492.

[109] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets”, in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[110] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang, “Weakly supervised disentangled generative causal representation learning”, *Journal of Machine Learning Research*, vol. 23, no. 241, 1, 2022.

[111] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Attribute and simile classifiers for face verification”, in *2009 IEEE 12th International Conference on Computer Vision*, 2009, 365. DOI: 10.1109/ICCV.2009.5459250.

[112] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer”, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 951. DOI: 10.1109/CVPR.2009.5206594.

[113] Z. Chen, Y. Bei, and C. Rudin, “Concept whitening for interpretable image recognition”, *Nature Machine Intelligence*, vol. 2, no. 12, 772, 2020. DOI: 10.1038/s42256-020-00265-z.

[114] D. Marcos, R. Fong, S. Lobry, R. Flamary, N. Courty, and D. Tuia, “Contextual semantic interpretability”, in *15th Asian Conference on Computer Vision*, H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, Eds., ser. Lecture Notes in Computer Science, vol. 12625, Springer, 2020, 351. DOI: 10.1007/978-3-030-69538-5_22.

[115] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)”, *IEEE Access*, vol. 6, 52138, 2018. DOI: 10.1109/ACCESS.2018.2870052.

[116] D. Gunning, E. Vorm, Y. Wang, and M. Turek, “DARPA’s explainable AI (XAI) program: A retrospective”, *Authorea*, 2021. DOI: 10.22541/au.163699841.19031727/v1.

[117] M. O. Karlsson and R. M. Savic, “Diagnosing model diagnostics”, *Clinical Pharmacology & Therapeutics*, vol. 82, no. 1, 17, 2007. DOI: 10.1038/sj.clpt.6100241.

[118] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks", *Genome Research*, vol. 28, no. 5, 739, 2018. DOI: `10.1101/gr.227819.117`.

[119] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients", *npj Digital Medicine*, vol. 2, no. 1, 2019. DOI: `10.1038/s41746-019-0172-3`.

[120] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks", in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, 3319.

[121] A. Chevan and M. Sutherland, "Hierarchical partitioning", *The American Statistician*, vol. 45, no. 2, 90, 1991. DOI: `10.1080/00031305.1991.10475776`.

[122] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach", *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, 319, 2001. DOI: `10.1002/asmb.446`.

[123] S. Hart, "Shapley value", in *Game Theory*, Palgrave Macmillan UK, 1989, 210. DOI: `10.1007/978-1-349-20181-5_25`.

[124] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[125] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees", *Nature Machine Intelligence*, vol. 2, no. 1, 56, 2020. DOI: `10.1038/s42256-019-0138-9`.

[126] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps", in *2nd International Conference on Learning Representations, ICLR 2014*, Y. Bengio and Y. LeCun, Eds., 2014.

[127] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences", in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, 3145.

[128] R. J. Aumann and L. S. Shapley, *Values of Non-Atomic Games*. Princeton: Princeton University Press, 1974. DOI: `10.1515/9781400867080`.

[129] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)", in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, 2668.

[130] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. A. Friedler, "Problems with Shapley-value-based explanations as feature importance measures", in *Proceedings of the 37th International Conference on Machine Learning*, JMLR.org, 2020.

[131] J. Adebayo, M. Muelly, H. Abelson, and B. Kim, "Post hoc explanations may be ineffective for detecting unknown spurious correlation", in *10th International Conference on Learning Representations, ICLR 2022*, OpenReview.net, 2022.

[132] A. M. Alaa and M. van der Schaar, "Demystifying black-box models with symbolic metamodels", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[133] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression", *Science Advances*, vol. 6, no. 16, eaay2631, 2020. DOI: `10.1126/sciadv.aay2631`.

[134] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019, 9277.

[135] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, 20554.

[136] J. L. Kolodner, "An introduction to case-based reasoning", *Artificial Intelligence Review*, vol. 6, no. 1, 3, 1992. DOI: `10.1007/bf00155578`.

[137] R. Caruana, H. Kangarloo, J. D. N. Dionisio, U. Sinha, and D. Johnson, "Case-based explanation of non-case-based learning methods", in *Proceedings of the AMIA Symposium*, 1999, 212.

[138] B. Kim, C. Rudin, and J. A. Shah, "The Bayesian case model: A generative approach for case-based reasoning and prototype classification", in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014, 1952.

[139] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability", in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016, 2288.

[140] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test", *Journal of Machine Learning Research*, vol. 13, no. 25, 723, 2012.

[141] J. Dodge, "Position: The case against case-based explanation.", in *Joint Proceedings of the ACM IUI Workshops 2022*, 2022, 175.

[142] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, Barcelona, Spain: Association for Computing Machinery, 2020, 607. DOI: `10.1145/3351095.3372850`.

[143] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 1765. DOI: `10.1109/CVPR.2017.17`.

[144] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation", in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[145] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning", in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019, 1. DOI: `10.1109/GlobalSIP45357.2019.8969491`.

[146] D. Mahajan, C. Tan, and A. Sharma, *Preserving causal constraints in counterfactual explanations for machine learning classifiers*, arXiv:1912.03277, 2019. DOI: 10.48550/arXiv.1912.03277.

[147] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare", *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 5, e1379, 2020. DOI: 10.1002/widm.1379.

[148] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use", in *Proceedings of the 4th Machine Learning for Healthcare Conference*, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., ser. Proceedings of Machine Learning Research, vol. 106, PMLR, 2019, 359.

[149] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care", *The Lancet Digital Health*, vol. 3, no. 11, e745, 2021. DOI: 10.1016/s2589-7500(21)00208-9.

[150] T. G. Dietterich, "Machine learning for sequential data: A review", in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2002, 15. DOI: 10.1007/3-540-70659-3_2.

[151] A. Allam, S. Feuerriegel, M. Rebhan, and M. Krauthammer, "Analyzing patient trajectories with artificial intelligence", *Journal of Medical Internet Research*, vol. 23, no. 12, e29812, 2021. DOI: 10.2196/29812.

[152] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica*, vol. 37, no. 3, 424, 1969. DOI: 10.2307/1912791.

[153] R. Marcinkevičs and J. E. Vogt, "Interpretable models for Granger causality using self-explaining neural networks", in *9th International Conference on Learning Representations, ICLR 2021*, OpenReview.net, 2021. DOI: 10.48550/arXiv.2101.07600.

[154] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems", in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, 2688.

[155] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[156] M. Eichler, "Causal inference in time series analysis", in *Causality*. John Wiley & Sons, Ltd, 2012, ch. 22, 327. DOI: 10.1002/9781119945710.ch22.

[157] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical Granger methods", in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07, San Jose, California, USA: Association for Computing Machinery, 2007, 66. DOI: 10.1145/1281192.1281203.

[158] W. B. Nicholson, D. S. Matteson, and J. Bien, "VARX-L: Structured regularization for large vector autoregressions with exogenous variables", *International Journal of Forecasting*, vol. 33, no. 3, 627, 2017. DOI: 10.1016/j.ijforecast.2017.01.003.

[159] L. Song, M. Kolar, and E. Xing, "Time-varying dynamic Bayesian networks", in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., vol. 22, Curran Associates, Inc., 2009.

[160] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, "Estimating time-varying networks", *Annals of Applied Statistics*, vol. 4, no. 1, 94, 2010. DOI: 10.1214/09-AOAS308.

[161] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality", *Physical Review Letters*, vol. 100, 144103, 14 2008. DOI: `10.1103/PhysRevLett.100.144103`.

[162] W. Ren, B. Li, and M. Han, "A novel Granger causality method based on HSIC-Lasso for revealing nonlinear relationship between multivariate time series", *Physica A: Statistical Mechanics and its Applications*, vol. 541, 123245, 2020. DOI: `10.1016/j.physa.2019.123245`.

[163] A. Montalto, S. Stramaglia, L. Faes, G. Tessitore, R. Prevete, and D. Marinazzo, "Neural networks with non-uniform embedding and explicit validation phase to assess Granger causality", *Neural Networks*, vol. 71, 159, 2015. DOI: `10.1016/j.neunet.2015.08.003`.

[164] Y. Wang, K. Lin, Y. Qi, Q. Lian, S. Feng, Z. Wu, and G. Pan, "Estimating brain connectivity with varying-length time lags using a recurrent neural network", *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, 1953, 2018. DOI: `10.1109/TBME.2018.2842769`.

[165] T. Wu, T. Breuel, M. Skuhersky, and J. Kautz, *Discovering nonlinear relations with minimum predictive information regularization*, arXiv:2001.01885, 2020. DOI: `10.48550/arXiv.2001.01885`.

[166] S. Löwe, D. Madras, R. Zemel, and M. Welling, "Amortized causal discovery: Learning to infer causal graphs from time-series data", in *Proceedings of the First Conference on Causal Learning and Reasoning*, B. Schölkopf, C. Uhler, and K. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 177, PMLR, 2022, 509.

[167] K. Inoue, A. Doncescu, and H. Nabeshima, "Hypothesizing about causal networks with positive and negative effects by meta-level abduction", in *Inductive Logic Programming*, Springer Berlin Heidelberg, 2011, 114. DOI: `10.1007/978-3-642-21295-6_15`.

[168] M. M. Rinschen, J. Ivanisevic, M. Giera, and G. Siuzdak, "Identification of bioactive metabolites using activity metabolomics", *Nature Reviews Molecular Cell Biology*, vol. 20, no. 6, 353, 2019. DOI: `10.1038/s41580-019-0108-4`.

[169] M. Pesaran, T. Schuermann, and S. M. Weiner, "Modeling regional interdependencies using a global error-correcting macroeconometric model", *Journal of Business & Economic Statistics*, vol. 22, no. 2, 129, 2004. DOI: `10.1198/073500104000000019`.

[170] S. Haufe, V. V. Nikulin, and G. Nolte, "Alleviating the influence of weak data asymmetries on Granger-causal analyses", in *Latent Variable Analysis and Signal Separation*, Springer Berlin Heidelberg, 2012, 25. DOI: `10.1007/978-3-642-28551-6_4`.

[171] I. Winkler, D. Panknin, D. Bartz, K.-R. Müller, and S. Haufe, "Validity of time reversal for testing Granger causality", *IEEE Transactions on Signal Processing*, vol. 64, no. 11, 2746, 2016. DOI: `10.1109/TSP.2016.2531628`.

[172] M. Chvosteková, J. Jakubík, and A. Krakovská, "Granger causality on forward and reversed time series", *Entropy*, vol. 23, no. 4, 409, 2021. DOI: `10.3390/e23040409`.

[173] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution", in *2010 20th International Conference on Pattern Recognition*, 2010, 3121. DOI: `10.1109/ICPR.2010.764`.

[174] L. A. Zager and G. C. Verghese, "Graph similarity scoring and matching", *Applied Mathematics Letters*, vol. 21, no. 1, 86, 2008. DOI: `10.1016/j.aml.2007.01.006`.

[175] N. Meinshausen and P. Bühlmann, "Stability selection", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, 417, 2010. DOI: `10.1111/j.1467-9868.2010.00740.x`.

[176] W. Sun, J. Wang, and Y. Fang, "Consistent selection of tuning parameters via variable selection stability", *Journal of Machine Learning Research*, vol. 14, no. 107, 3419, 2013.

[177] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, vol. 9, no. 8, 1735, 1997. DOI: `10.1162/neco.1997.9.8.1735`.

[178] Y. Chen, Y. Kang, Y. Chen, and Z. Wang, "Probabilistic forecasting with temporal convolutional neural network", *Neurocomputing*, vol. 399, 491, 2020. DOI: `10.1016/j.neucom.2020.03.011`.

[179] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, *A dual-stage attention-based recurrent neural network for time series prediction*, arXiv:1704.02971, 2017. DOI: `10.48550/arXiv.1704.02971`.

[180] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, 289, 1995. DOI: `10.1111/j.2517-6161.1995.tb02031.x`.

[181] E. N. Lorenz, "Predictability: A problem partly solved", in *Seminar on Predictability*, vol. 1, Shinfield Park, Reading: ECMWF, 1995, 1.

[182] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for FMRI", *NeuroImage*, vol. 54, no. 2, 875, 2011. DOI: `10.1016/j.neuroimage.2010.08.063`.

[183] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using Granger causality and fMRI", *NeuroImage*, vol. 25, no. 1, 230, 2005. DOI: `10.1016/j.neuroimage.2004.11.017`.

[184] N. Bacaër, "Lotka, Volterra and the predator–prey system (1920–1926)", in *A Short History of Mathematical Population Dynamics*. London: Springer London, 2011, 71. DOI: `10.1007/978-0-85729-115-8_13`.

[185] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves", in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, 233. DOI: `10.1145/1143844.1143874`.

[186] N. Parikh and S. Boyd, "Proximal algorithms", *Foundations and Trends® in Optimization*, vol. 1, no. 3, 127, 2014. DOI: `10.1561/2400000003`.

[187] B. Bussmann, J. Nys, and S. Latré, "Neural additive vector autoregression models for causal discovery in time series", in *Discovery Science*, C. Soares and L. Torgo, Eds., Cham: Springer International Publishing, 2021, 446. DOI: `10.1007/978-3-030-88942-5_35`.

[188] D. K. Christopoulos and M. A. León-Ledesma, "Testing for Granger (non-)causality in a time-varying coefficient VAR model", *Journal of Forecasting*, vol. 27, no. 4, 293, 2008. DOI: `10.1002/for.1060`.

[189] G. Koop and D. Korobilis, "Large time-varying parameter VARs", *Journal of Econometrics*, vol. 177, no. 2, 185, 2013. DOI: `10.1016/j.jeconom.2013.04.007`.

[190] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, "Network physiology reveals relations between network topology and physiological function", *Nature Communications*, vol. 3, no. 1, 2012. DOI: `10.1038/ncomms1705`.

[191] R. Marcinkevics, "Causal inference in time series for identifying molecular fingerprints during sleep", M.S. thesis, ETH Zurich, 2019.

[192] N. Nowak, T. Gaisl, D. Miladinovic, R. Marcinkevics, M. Osswald, S. Bauer, J. Buhmann, R. Zenobi, P. Sinues, S. A. Brown, and M. Kohler, "Rapid and reversible control of human metabolism by individual sleep states", *Cell Reports*, vol. 37, no. 4, 109903, 2021. DOI: `10.1016/j.celrep.2021.109903`.

[193] A. H. Hatteland, R. Marcinkevičs, R. Marquis, T. Frick, I. Hubbard, J. E. Vogt, T. Brunschwiler, and P. Ryvlin, "Exploring relationships between cerebral and peripheral biosignals with neural networks", in *2021 IEEE International Conference on Digital Health (ICDH)*, 2021, 103. DOI: `10.1109/ICDH52753.2021.00022`.

[194] Z. Xiao, M. Muszynski, R. Marcinkevičs, L. Zimmerli, A. D. Ivankay, D. Kohlbrenner, M. Kuhn, Y. Nordmann, U. Muehlner, C. Clarenbach, J. E. Vogt, and T. Brunschwiler, "Breathing new life into COPD assessment: Multisensory home-monitoring for predicting severity", in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23, Paris, France: Association for Computing Machinery, 2023, 84. DOI: `10.1145/3577190.3614109`.

[195] O. Lawal, W. M. Ahmed, T. M. E. Nijsen, R. Goodacre, and S. J. Fowler, "Exhaled breath analysis: A review of 'breath-taking' methods for off-line analysis", *Metabolomics*, vol. 13, no. 10, 2017. DOI: `10.1007/s11306-017-1241-8`.

[196] C. L. Scrivener and A. T. Reader, "Variability of EEG electrode positions and their underlying brain regions: Visualizing gel artifacts from a simultaneous EEG-fMRI dataset", *Brain and Behavior*, vol. 12, no. 2, e2476, 2022. DOI: `10.1002/brb3.2476`.

[197] S. C.-X. Li and B. Marlin, "Learning from irregularly-sampled time series: A missing data perspective", in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, 5937.

[198] D. Kohlbrenner, C. F. Clarenbach, A. Ivankay, L. Zimmerli, C. S. Gross, M. Kuhn, and T. Brunschwiler, "Multisensory home-monitoring in individuals with stable chronic obstructive pulmonary disease and asthma: Usability study of the CAir-Desk", *JMIR Human Factors*, vol. 9, no. 1, e31448, 2022. DOI: `10.2196/31448`.

[199] M. Maziarz, "A review of the Granger-causality fallacy", *The Journal of Philosophical Economics*, vol. 8, no. 2, 6, 2015.

[200] N. Jethani, M. Sudarshan, Y. Aphinyanaphongs, and R. Ranganath, "Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations", in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, 1459.

[201] A. Madar, A. Greenfield, E. Vanden-Eijnden, and R. Bonneau, "DREAM3: Network inference using dynamic context likelihood of relatedness and the inferelator", *PLOS ONE*, vol. 5, no. 3, 1, 2010. DOI: `10.1371/journal.pone.0009803`.

[202] S. Cekic, D. Grandjean, and O. Renaud, "Time, frequency, and time-varying Granger-causality measures in neuroscience", *Statistics in Medicine*, vol. 37, no. 11, 1910, 2018. DOI: `10.1002/sim.7621`.

[203] H. Xu, M. Farajtabar, and H. Zha, "Learning Granger causality for Hawkes processes", in *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan and K. Q. Weinberger, Eds., ser. Proceedings of Machine Learning Research, vol. 48, New York, New York, USA: PMLR, 2016, 1717.

[204] A. Tank, X. Li, E. B. Fox, and A. Shojaie, "The convex mixture distribution: Granger causality for categorical time series", *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 1, 83, 2021. DOI: `10.1137/20M133097X`.

[205] Đ. Miladinović, C. Muheim, S. Bauer, A. Spinnler, D. Noain, M. Bandarabadi, B. Gallusser, G. Krummenacher, C. Baumann, A. Adamantidis, S. A. Brown, and J. M. Buhmann, "SPINDLE: End-to-end learning from EEG/EMG to extrapolate animal sleep scoring across experimental settings, labs and species", *PLOS Computational Biology*, vol. 15, no. 4, 1, 2019. DOI: `10.1371/journal.pcbi.1006968`.

[206] G. Rodríguez, *Lecture notes on generalized linear models*, 2007.

[207] D. G. Altman, "Analysis of survival times", in *Practical Statistics for Medical Research*, Chapman and Hall/CRC Texts in Statistical Science Series, 2020, ch. 13, 365. DOI: `10.1201/9780429258589`.

[208] D. Faraggi and R. Simon, "A neural network model for survival data", *Statistics in Medicine*, vol. 14, no. 1, 73, 1995. DOI: `10.1002/sim.4780140108`.

[209] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei, "Deep survival analysis", in *Proceedings of the 1st Machine Learning for Healthcare Conference*, F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, Eds., ser. Proceedings of Machine Learning Research, vol. 56, Northeastern University, Boston, MA, USA: PMLR, 2016, 101.

[210] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network", *BMC Medical Research Methodology*, vol. 18, no. 1, 2018. DOI: `10.1186/s12874-018-0482-1`.

[211] C. Lee, W. Zame, J. Yoon, and M. van der Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. DOI: `10.1609/aaai.v32i1.11842`.

[212] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and Cox regression", *Journal of Machine Learning Research*, vol. 20, no. 129, 1, 2019.

[213] C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, 3163, 2021. DOI: `10.1109/JBHI.2021.3052441`.

[214] L. Manduchi, R. Marcinkevičs, M. C. Massi, T. J. Weikert, A. Sauter, V. Gotta, T. Müller, F. Vasella, M. C. Neidert, M. Pfister, B. Stieltjes, and J. E. Vogt, "A deep variational approach to clustering survival data", in *10th International Conference on Learning Representations, ICLR 2022*, OpenReview.net, 2022. DOI: `10.48550/arXiv.2106.05763`.

[215] M. Winkel, *Statistical lifetime-models*, Lecture notes. University of Oxford, 2007.

[216] D. R. Cox, "Regression models and life-tables", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, 187, 1972. DOI: `10.1111/j.2517-6161.1972.tb00899.x`.

[217] D. R. Cox and D. Oakes, *Analysis of Survival Data*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1984.

[218] D. R. Cox, "Partial likelihood", *Biometrika*, vol. 62, no. 2, 269, 1975. DOI: `10.1093/biomet/62.2.269`.

[219] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests", *The Annals of Applied Statistics*, vol. 2, no. 3, 841, 2008. DOI: `10.1214/08-AOAS169`.

[220] R. Ranganath, L. Tang, L. Charlin, and D. Blei, "Deep exponential families", in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, G. Lebanon and S. V. N. Vishwanathan, Eds., ser. Proceedings of Machine Learning Research, vol. 38, San Diego, California, USA: PMLR, 2015, 762.

[221] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spégel, H. Mulder, E. Lindholm, O. Melander, O. Hansson, U. Malmqvist, Å. Lernmark, K. Lahti, T. Forsén, T. Tuomi, A. H. Rosengren, and L. Groop, "Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables", *The Lancet Diabetes & Endocrinology*, vol. 6, no. 5, 361, 2018. DOI: `10.1016/s2213-8587(18)30051-2`.

[222] F. S. Collins and H. Varmus, "A new initiative on precision medicine", *New England Journal of Medicine*, vol. 372, no. 9, 793, 2015. DOI: `10.1056/nejmp1500523`.

[223] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data", *PLoS Biology*, vol. 2, no. 4, T. Golub, Ed., e108, 2004. DOI: `10.1371/journal.pbio.0020108`.

[224] D. A. Fenstermacher, R. M. Wenham, D. E. Rollison, and W. S. Dalton, "Implementing personalized medicine in a cancer center", *The Cancer Journal*, vol. 17, no. 6, 528, 2011. DOI: `10.1097/ppo.0b013e318238216e`.

[225] V. T. Farewell, "The use of mixture models for the analysis of survival data with long-term survivors", *Biometrics*, vol. 38, no. 4, 1041, 1982. DOI: `10.2307/2529885`.

[226] S. Mouli, B. Ribeiro, and J. Neville, *A deep learning approach for survival clustering without end-of-life signals*, 2018.

[227] E. Xia, X. Du, J. Mei, W. Sun, S. Tong, Z. Kang, J. Sheng, J. Li, C. Ma, J. Dong, and S. Li, *Outcome-driven clustering of acute coronary syndrome patients using multi-task neural network with attention*, arXiv:1903.00197, 2019. DOI: `10.48550/arXiv.1903.00197`.

[228] P. Chapfuwa, C. Li, N. Mehta, L. Carin, and R. Henao, "Survival cluster analysis", in *Proceedings of the ACM Conference on Health, Inference, and Learning*, Toronto, Ontario, Canada: Association for Computing Machinery, 2020, 60. DOI: `10.1145/3368555.3384465`.

[229] C. Nagpal, S. Yadlowsky, N. Rostamzadeh, and K. Heller, "Deep Cox mixtures for survival regression", in *Proceedings of the 6th Machine Learning for Healthcare Conference*, K. Jung, S. Yeung, M. Sendak, M. Sjoding, and R. Ranganath, Eds., ser. Proceedings of Machine Learning Research, vol. 149, PMLR, 2021, 674.

[230] S. Liverani, L. Leigh, I. L. Hudson, and J. E. Byles, "Clustering method for censored and collinear survival data", *Computational Statistics*, vol. 36, no. 1, 35, 2020. DOI: `10.1007/s00180-020-01000-3`.

[231] J. Molitor, M. Papathomas, M. Jerrett, and S. Richardson, "Bayesian profile regression with an application to the National survey of children's health", *Biostatistics*, vol. 11, no. 3, 484, 2010. DOI: `10.1093/biostatistics/kxq013`.

[232] C. M. Bishop, "Latent variable models", in *Learning in Graphical Models*. Springer Netherlands, 1998, 371. DOI: `10.1007/978-94-011-5014-9_13`.

[233] D. P. Kingma and M. Welling, “An introduction to variational autoencoders”, *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, 307, 2019. DOI: 10.1561/2200000056.

[234] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, *Deep unsupervised clustering with Gaussian mixture variational autoencoders*, arXiv:1611.02648, 2016. DOI: 10.48550/arXiv.1611.02648.

[235] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, “Variational deep embedding: An unsupervised and generative approach to clustering”, in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI’17, Melbourne, Australia: AAAI Press, 2017, 1965.

[236] A. Skrondal and S. Rabe-Hesketh, “Latent variable modelling: A survey”, *Scandinavian Journal of Statistics*, vol. 34, no. 4, 712, 2007. DOI: 10.1111/j.1467-9469.2007.00573.x.

[237] P.-A. Mattei and J. Frellsen, “Leveraging the exact likelihood of deep latent variable models”, in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018.

[238] S. J. Gershman and N. D. Goodman, “Amortized inference in probabilistic reasoning”, *Cognitive Science*, vol. 36, 2014.

[239] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models”, in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 2014, 1278.

[240] A. B. Dieng, *Deep probabilistic graphical modeling*, arXiv:2104.12053, 2021. DOI: 10.48550/arXiv.2104.12053.

[241] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, “Semi-supervised learning with deep generative models”, in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014.

[242] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models”, in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015.

[243] M. Wu and N. Goodman, “Multimodal generative models for scalable weakly-supervised learning”, in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018.

[244] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, “Finite mixture models”, *Annual Review of Statistics and Its Application*, vol. 6, no. 1, 355, 2019. DOI: 10.1146/annurev-statistics-031017-100325.

[245] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts”, *Neural Computation*, vol. 3, no. 1, 79, 1991. DOI: 10.1162/NECO.1991.3.1.79.

[246] I. C. Gormley and S. Frühwirth-Schnatter, *Mixtures of experts models*, arXiv:1806.08200, 2018. DOI: 10.48550/arXiv.1806.08200.

[247] S. Gaffney and P. Smyth, “Trajectory clustering with mixtures of regression models”, in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’99, San Diego, California, USA: Association for Computing Machinery, 1999, 63. DOI: 10.1145/312129.312198.

[248] M. I. Jordan, *An introduction to graphical models*, 1997.

[249] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, 1, 1977. DOI: `10.1111/j.2517-6161.1977.tb01600.x`.

[250] K. J. Carroll, "On the use and utility of the Weibull model in the analysis of survival data", *Controlled Clinical Trials*, vol. 24, no. 6, 682, 2003. DOI: `10.1016/S0197-2456(03)00072-2`.

[251] S. Pölsterl, *Survival analysis for deep learning*, `https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/`, 2019.

[252] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner, "The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults", *Annals of Internal Medicine*, vol. 122, no. 3, 191, 1995. DOI: `10.7326/0003-4819-122-3-199502010-00007`.

[253] J. C. Buckner, "Factors influencing survival in high-grade gliomas", *Seminars in Oncology*, vol. 30, 10, 2003. DOI: `10.1053/j.seminoncol.2003.11.031`.

[254] M. Weller, M. van den Bent, M. Preusser, E. Le Rhun, J. C. Tonn, G. Minniti, M. Bendszus, C. Balana, O. Chinot, L. Dirven, P. French, M. E. Hegi, A. S. Jakola, M. Platten, P. Roth, R. Rudà, S. Short, M. Smits, M. J. B. Taphoorn, A. von Deimling, M. Westphal, R. Soffietti, G. Reifenberger, and W. Wick, "EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood", *Nature Reviews Clinical Oncology*, vol. 18, no. 3, 170, 2020. DOI: `10.1038/s41571-020-00447-z`.

[255] V. Gotta, O. Marsenic, and M. Pfister, "Age- and weight-based differences in haemodialysis prescription and delivery in children, adolescents and young adults", *Nephrology Dialysis Transplantation*, vol. 33, no. 9, 1649, 2018. DOI: `10.1093/ndt/gfy067`.

[256] V. Gotta, M. Pfister, and O. Marsenic, "Ultrafiltration rates in children on chronic hemodialysis routinely exceed weight based adult limit", *Hemodialysis International*, vol. 23, no. 1, 126, 2019. DOI: `10.1111/hdi.12727`.

[257] V. Gotta, O. Marsenic, and M. Pfister, "Understanding urea kinetic factors that enhance personalized hemodialysis prescription in children", *ASAIO Journal*, vol. 66, no. 1, 115, 2020. DOI: `10.1097/mat.0000000000000941`.

[258] V. Gotta, G. Tancev, O. Marsenic, J. E. Vogt, and M. Pfister, "Identifying key predictors of mortality in young patients on chronic haemodialysis—–-a machine learning approach", *Nephrology Dialysis Transplantation*, vol. 36, no. 3, 519, 2020. DOI: `10.1093/ndt/gfaa128`.

[259] V. Gotta, O. Marsenic, A. Atkinson, and M. Pfister, "Hemodialysis (HD) dose and ultrafiltration rate are associated with survival in pediatric and adolescent patients on chronic HD—a large observational study with follow-up to young adult age", *Pediatric Nephrology*, 2021. DOI: `10.1007/s00467-021-04972-6`.

[260] T. Weikert, T. Akinci D'Antonoli, J. Bremerich, B. Stieltjes, G. Sommer, and A. W. Sauter, "Evaluation of an AI-powered lung nodule algorithm for detection and 3D segmentation of primary lung tumors", *Contrast Media & Molecular Imaging*, vol. 2019, 1, 2019. DOI: `10.1155/2019/1545747`.

[261] H. J. W. L. Aerts, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach", *Nature Communications*, vol. 5, no. 1, 2014. DOI: `10.1038/ncomms5006`.

[262] H. J. W. L. Aerts, L. Wee, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, *Data from NSCLC-radiomics*, 2019. DOI: `10.7937/K9/TCIA.2015.PF0M9REI`.

[263] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository", *Journal of Digital Imaging*, vol. 26, no. 6, 1045, 2013. DOI: `10.1007/s10278-013-9622-7`.

[264] H. J. W. L. Aerts, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, *Data from NSCLC-radiomics-genomics*, 2015. DOI: `10.7937/K9/TCIA.2015.L4FRET6Z`.

[265] O. Gevaert, J. Xu, C. D. Hoang, A. N. Leung, Y. Xu, A. Quon, D. L. Rubin, S. Napel, and S. K. Plevritis, "Non–small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results", *Radiology*, vol. 264, no. 2, 387, 2012. DOI: `10.1148/radiol.12111607`.

[266] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, W. Zhang, A. Leung, M. Kadoch, J. Shrager, A. Quon, D. Rubin, S. Plevritis, and S. Napel, *Data for NSCLC radiogenomics collection*, 2017. DOI: `10.7937/K9/TCIA.2017.7hs46erv`.

[267] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, A. N. C. Leung, M. Kadoch, C. D. Hoang, J. Shrager, A. Quon, D. L. Rubin, S. K. Plevritis, and S. Napel, "A radiogenomic dataset of non-small cell lung cancer", *Scientific Data*, vol. 5, no. 1, 2018. DOI: `10.1038/sdata.2018.202`.

[268] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness", in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, 29935.

[269] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype", *Cancer Research*, vol. 77, no. 21, e104, 2017. DOI: `10.1158/0008-5472.CAN-17-0339`.

[270] V. C. Raykar, H. Steck, B. Krishnapuram, C. Dehing-Oberije, and P. Lambin, "On ranking in survival analysis: Bounds on the concordance index", in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, Curran Associates, Inc., 2007, 1209.

[271] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors", in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24, Curran Associates, Inc., 2011, 1845.

[272] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics", *BMC Medicine*, vol. 17, no. 1, 2019. DOI: 10.1186/s12916-019-1466-7.

[273] M. Goldstein, X. Han, A. Puli, A. Perotte, and R. Ranganath, "X-CAL: Explicit calibration for survival analysis", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, 18296.

[274] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations", *Journal of the American Statistical Association*, vol. 53, no. 282, 457, 1958. DOI: 10.1080/01621459.1958.10501452.

[275] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE", *Journal of Machine Learning Research*, vol. 9, no. 86, 2579, 2008.

[276] H. W. Lee, C.-H. Lee, and Y. S. Park, "Location of stage I–III non-small cell lung cancer and survival rate: Systematic review and meta-analysis", *Thoracic Cancer*, vol. 9, no. 12, 1614, 2018. DOI: 10.1111/1759-7714.12869.

[277] D. Etiz, L. B. Marks, S.-M. Zhou, G. C. Bentel, R. Clough, M. L. Hernando, and P. A. Lind, "Influence of tumor volume on survival in patients irradiated for non-small-cell lung cancer", *International Journal of Radiation Oncology*Biology*Physics*, vol. 53, no. 4, 835, 2002. DOI: 10.1016/S0360-3016(02)02814-6.

[278] M. Agarwal, G. Brahmanday, G. W. Chmielewski, R. J. Welsh, and K. P. Ravikrishnan, "Age, tumor size, type of surgery, and gender predict survival in early stage (stage I and II) non-small cell lung cancer after surgical resection", *Lung Cancer*, vol. 68, no. 3, 398, DOI: 10.1016/j.lungcan.2009.08.008.

[279] V. R. Bhatt, R. Batra, P. T. Silberstein, F. R. Loberiza, and A. K. Ganti, "Effect of smoking on survival from non-small cell lung cancer: A retrospective Veterans' Affairs Central Cancer Registry (VACCR) cohort analysis", *Medical Oncology*, vol. 32, no. 1, 2014. DOI: 10.1007/s12032-014-0339-3.

[280] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis", *Journal of the American Statistical Association*, vol. 47, no. 260, 583, 1952. DOI: 10.1080/01621459.1952.10483441.

[281] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-based validation of clustering solutions", *Neural Computation*, vol. 16, no. 6, 1299, 2004. DOI: 10.1162/089976604773717621.

[282] C. Haarburger, P. Weitz, O. Rippel, and D. Merhof, "Image-based survival prediction for lung cancer patients using CNNs", in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019. DOI: 10.1109/isbi.2019.8759499.

[283] G. A. Bello, T. J. W. Dawes, J. Duan, C. Biffi, A. de Marvao, L. S. G. E. Howard, J. S. R. Gibbs, M. R. Wilkins, S. A. Cook, D. Rueckert, and D. P. O'Regan, "Deep-learning cardiac motion analysis for human survival prediction", *Nature Machine Intelligence*, vol. 1, no. 2, 95, 2019. DOI: 10.1038/s42256-019-0019-2.

[284] C. Lee, J. Yoon, and M. van der Schaar, "Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data", *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, 122, 2020. DOI: 10.1109/tbme.2019.2909027.

[285] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, *Towards a definition of disentangled representations*, arXiv:1812.02230, 2018. DOI: 10.48550/arXiv.1812.02230.

[286] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models", in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, 21696.

[287] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, 19667.

[288] Đ. Miladinovic, A. Stanic, S. Bauer, J. Schmidhuber, and J. M. Buhmann, "Spatial dependency networks: Neural layers for improved generative image modeling", in *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[289] Y. Shi, S. N, B. Paige, and P. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019, 15718.

[290] L. M. Wier, H. Yu, P. L. Owens, and R. Washington, *Overview of Children in the Emergency Department, 2010: Statistical Brief #157*. Rockville, MD, USA: Agency for Healthcare Research and Quality, 2013.

[291] E. Kobayashi, B. Johnson, K. Goetz, J. Scanlan, and R. Weinsheimer, "Does the implementation of a pediatric appendicitis pathway promoting ultrasound work outside of a children's hospital?", *The American Journal of Surgery*, vol. 215, no. 5, 917, 2018, North Pacific Surgical Association. DOI: 10.1016/j.amjsurg.2018.03.017.

[292] R. Marcinkevičs, P. Reis Wolfertstetter, U. Klimiene, K. Chin-Cheong, A. Paschke, J. Zerres, M. Denzinger, D. Niederberger, S. Wellmann, E. Ozkan, C. Knorr, and J. E. Vogt, "Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis", *Medical Image Analysis*, vol. 91, 103042, 2024. DOI: 10.1016/j.media.2023.103042.

[293] G. Alain and Y. Bengio, *Understanding intermediate layers using linear classifier probes*, arXiv:1610.01644, 2016. DOI: 10.48550/arXiv.1610.01644.

[294] Y. Belinkov, "Probing classifiers: Promises, shortcomings, and advances", *Computational Linguistics*, vol. 48, no. 1, 207, 2022. DOI: 10.1162/coli_a_00422.

[295] R. Caruana, "Multitask learning", *Machine Learning*, vol. 28, no. 1, 41, 1997. DOI: 10.1023/a:1007379606734.

[296] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, "Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning", *Ultrasound in Medicine & Biology*, vol. 46, no. 5, 1119, 2020. DOI: 10.1016/j.ultrasmedbio.2020.01.001.

[297] X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng, Q. Sun, L. Lu, and K. K. Shung, "Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning", *Nature Biomedical Engineering*, vol. 5, no. 6, 522, 2021. DOI: 10.1038/s41551-021-00711-2.

[298] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, 6827.

[299] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, "Self-supervised learning with data augmentations provably isolates content from style", in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, 16451.

[300] M. Suzuki and Y. Matsuo, "A survey of multimodal deep generative models", *Advanced Robotics*, vol. 36, no. 5-6, 261, 2022. DOI: `10.1080/01691864.2022.2035253`.

[301] S. Di Saverio, A. Birindelli, M. D. Kelly, F. Catena, D. G. Weber, M. Sartelli, M. Sugrue, M. De Moya, C. A. Gomes, A. Bhangu, F. Agresta, E. E. Moore, K. Soreide, E. Griffiths, S. De Castro, J. Kashuk, Y. Kluger, A. Leppaniemi, L. Ansaloni, M. Andersson, F. Coccolini, R. Coimbra, K. S. Gurusamy, F. C. Campanile, W. Biffl, O. Chiara, F. Moore, A. B. Peitzman, G. P. Fraga, D. Costa, R. V. Maier, S. Rizoli, Z. J. Balogh, C. Bendinelli, R. Cirocchi, V. Tonini, A. Piccinini, G. Tugnoli, E. Jovine, R. Persiani, A. Biondi, T. Scalea, P. Stahel, R. Ivatury, G. Velmahos, and R. Andersson, "WSES Jerusalem guidelines for diagnosis and treatment of acute appendicitis", *World Journal of Emergency Surgery*, vol. 11, no. 1, 2016. DOI: `10.1186/s13017-016-0090-5`.

[302] A. Acharya, S. R. Markar, M. Ni, and G. B. Hanna, "Biomarkers of acute appendicitis: Systematic review and cost–benefit trade-off analysis", *Surgical Endoscopy*, vol. 31, no. 3, 1022, 2016. DOI: `10.1007/s00464-016-5109-1`.

[303] N. Kiss, M. Minderjahn, J. Reismann, J. Svensson, T. Wester, K. Hauptmann, M. Schad, J. Kallarackal, H. von Bernuth, and M. Reismann, "Use of gene expression profiling to identify candidate genes for pretherapeutic patient classification in acute appendicitis", *BJS Open*, vol. 5, no. 1, zraa045, 2021. DOI: `10.1093/bjsopen/zraa045`.

[304] R. Andersson, A. Hugander, A. Thulin, P. O. Nystrom, and G. Olaison, "Indications for operation in suspected appendicitis and incidence of perforation", *BMJ*, vol. 308, no. 6921, 107, 1994. DOI: `10.1136/bmj.308.6921.107`.

[305] A. Bhangu, K. Søreide, S. Di Saverio, J. H. Assarsson, and F. T. Drake, "Acute appendicitis: Modern understanding of pathogenesis, diagnosis, and management", *The Lancet*, vol. 386, no. 10000, 1278, 2015. DOI: `10.1016/S0140-6736(15)00275-5`.

[306] R. R. Gorter, H. H. Eker, M. A. W. Gorter-Stam, G. S. A. Abis, A. Acharya, M. Ankersmit, S. A. Antoniou, S. Arolfo, B. Babic, L. Boni, M. Bruntink, D. A. van Dam, B. Defoort, C. L. Deijen, F. B. DeLacy, P. M. Go, A. M. K. Harmsen, R. S. van den Helder, F. Iordache, J. C. F. Ket, F. E. Muysoms, M. M. Ozmen, M. Papoulas, M. Rhodes, J. Straatman, M. Tenhagen, V. Turrado, A. Vereczkei, R. Vilallonga, J. D. Deelder, and J. Bonjer, "Diagnosis and management of acute appendicitis. EAES consensus development conference 2015", *Surgical Endoscopy*, vol. 30, no. 11, 4668, 2016. DOI: `10.1007/s00464-016-5245-7`.

[307] J. Svensson, N. Hall, S. Eaton, A. Pierro, and T. Wester, "A review of conservative treatment of acute appendicitis", *European Journal of Pediatric Surgery*, vol. 22, no. 03, 185, 2012. DOI: `10.1055/s-0032-1320014`.

[308] J. F. Svensson, B. Patkova, M. Almström, H. Naji, N. J. Hall, S. Eaton, A. Pierro, and T. Wester, "Nonoperative treatment with antibiotics versus surgery for acute nonperforated appendicitis in children", *Annals of Surgery*, vol. 261, no. 1, 67, 2015. DOI: 10.1097/sla.0000000000000835.

[309] CODA Collaborative, "A randomized trial comparing antibiotics with appendectomy for appendicitis", *New England Journal of Medicine*, vol. 383, no. 20, 1907, 2020. DOI: 10.1056/nejmoa2014320.

[310] N. H. Park, H. E. Oh, H. J. Park, and J. Y. Park, "Ultrasonography of normal and abnormal appendix in children", *World Journal of Radiology*, vol. 3, no. 4, 85, 2011. DOI: 10.4329/wjr.v3.i4.85.

[311] J. Dingemann and B. Ure, "Imaging and the use of scores for the diagnosis of appendicitis in children", *European Journal of Pediatric Surgery*, vol. 22, no. 03, 195, 2012. DOI: 10.1055/s-0032-1320017.

[312] G. Ohba, S. Hirobe, and K. Komori, "The usefulness of combined B mode and Doppler ultrasonography to guide treatment of appendicitis", *European Journal of Pediatric Surgery*, vol. 26, no. 06, 533, 2016. DOI: 10.1055/s-0035-1570756.

[313] C.-H. Hsieh, R.-H. Lu, N.-H. Lee, W.-T. Chiu, M.-H. Hsu, and Y.-C. Li, "Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks", *Surgery*, vol. 149, no. 1, 87, 2011. DOI: 10.1016/j.surg.2010.03.023.

[314] L. Deleger, H. Brodzinski, H. Zhai, Q. Li, T. Lingren, E. S. Kirkendall, E. Alessandrini, and I. Solti, "Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department", *Journal of the American Medical Informatics Association*, vol. 20, no. e2, e212, 2013. DOI: 10.1136/amiajnl-2013-001962.

[315] J. Reismann, A. Romualdi, N. Kiss, M. I. Minderjahn, J. Kallarackal, M. Schad, and M. Reismann, "Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach", *PLoS ONE*, vol. 14, no. 9, e0222030, 2019. DOI: 10.1371/journal.pone.0222030.

[316] E. Aydin, İ. U. Türkmen, G. Namli, Ç. Öztürk, A. B. Esen, Y. N. Eray, E. Eroglu, and F. Akova, "A novel and simple machine learning algorithm for preoperative diagnosis of acute appendicitis in children", *Pediatric Surgery International*, vol. 36, no. 6, 735, 2020. DOI: 10.1007/s00383-020-04655-7.

[317] O. F. Akmese, G. Dogan, H. Kor, H. Erbay, and E. Demir, "The use of machine learning approaches for the diagnosis of acute appendicitis", *Emergency Medicine International*, vol. 2020, 1, 2020. DOI: 10.1155/2020/7306435.

[318] C. Stiel, J. Elrod, M. Klinke, J. Herrmann, C.-M. Junge, T. Ghadban, K. Reinshagen, and M. Boettcher, "The modified Heidelberg and the AI appendicitis score are superior to current scores in predicting appendicitis in children: A two-center cohort study", *Frontiers in Pediatrics*, vol. 8, 2020. DOI: 10.3389/fped.2020.592892.

[319] P. Rajpurkar, A. Park, J. Irvin, C. Chute, M. Bereket, D. Mastrodicasa, C. P. Langlotz, M. P. Lungren, A. Y. Ng, and B. N. Patel, "AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining", *Scientific Reports*, vol. 10, no. 1, 2020. DOI: 10.1038/s41598-020-61055-6.

[320] R. Marcinkevics, P. Reis Wolfertstetter, S. Wellmann, C. Knorr, and J. E. Vogt, "Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis", *Frontiers in Pediatrics*, vol. 9, 2021. DOI: `10.3389/fped.2021.662183`.

[321] P. Roig Aparicio, R. Marcinkevičs, P. Reis Wolfertstetter, S. Wellmann, C. Knorr, and J. E. Vogt, "Learning medical risk scores for pediatric appendicitis", in *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pasadena, CA, USA: IEEE, 2021. DOI: `10.1109/ICMLA52953.2021.00243`.

[322] J. Xia, Z. Wang, D. Yang, R. Li, G. Liang, H. Chen, A. A. Heidari, H. Turabieh, M. Mafarja, and Z. Pan, "Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis", *Computers in Biology and Medicine*, vol. 143, 105206, 2022. DOI: `10.1016/j.compbiomed.2021.105206`.

[323] G. Mostbeck, E. J. Adam, M. B. Nielsen, M. Claudon, D. Clevert, C. Nicolau, C. Nyhsen, and C. M. Owens, "How to diagnose acute appendicitis: Ultrasound first", *Insights into Imaging*, vol. 7, no. 2, 255, 2016. DOI: `10.1007/s13244-016-0469-6`.

[324] V. Cheplygina, "Cats or CAT scans: Transfer learning from natural or medical image source data sets?", *Current Opinion in Biomedical Engineering*, vol. 9, 21, 2019. DOI: `10.1016/j.cobme.2018.12.005`.

[325] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Athens, Greece: Springer International Publishing, 2016, 469. DOI: `10.1007/978-3-319-46723-8_54`.

[326] C. Ma, Y. Guo, J. Yang, and W. An, "Learning multi-view representation with LSTM for 3-D shape recognition and retrieval", *IEEE Transactions on Multimedia*, vol. 21, no. 5, 1169, 2019. DOI: `10.1109/TMM.2018.2875512`.

[327] E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles, and K. M. Pohl, "Representation learning with statistical independence to mitigate bias", in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, 2021. DOI: `10.1109/wacv48630.2021.00256`.

[328] M. Cogswell, F. Ahmed, R. B. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations", in *4th International Conference on Learning Representations, ICLR 2016*, Y. Bengio and Y. LeCun, Eds., OpenReview.net, 2016.

[329] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks", *Communications of the ACM*, vol. 63, no. 11, 139, 2020. DOI: `10.1145/3422622`.

[330] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *Caltech-UCSD Birds-200-2011*, Technical report. CNS-TR-2011-001. California Institute of Technology. `https://authors.library.caltech.edu/records/cvm3y-5hh21`, 2011.

[331] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, 2251, 2019. DOI: `10.1109/tpami.2018.2857768`.

[332] R. Marcinkevičs, P. Reis Wolfertstetter, U. Klimiene, K. Chin-Cheong, A. Paschke, J. Zerres, M. Denzinger, D. Niederberger, S. Wellmann, E. Ozkan, C. Knorr, and J. E. Vogt, *Regensburg pediatric appendicitis dataset*, version 1.03, Zenodo, 2023. DOI: `10.5281/zenodo.7711412`.

[333] A. Alvarado, “A practical score for the early diagnosis of acute appendicitis”, *Annals of Emergency Medicine*, vol. 15, no. 5, 557, 1986. DOI: `10.1016/S0196-0644(86)80993-3`.

[334] M. Samuel, “Pediatric appendicitis score”, *Journal of Pediatric Surgery*, vol. 37, no. 6, 877, 2002. DOI: `10.1053/jpsu.2002.32893`.

[335] Collaborative RSGobotWMR, “Appendicitis risk prediction models in children presenting with right iliac fossa pain (RIFT study): A prospective, multicentre validation study”, *The Lancet Child & Adolescent Health*, vol. 4, no. 4, 271, 2020. DOI: `10.1016/S2352-4642(20)30006-7`.

[336] J. Dingemann and B. Ure, “Imaging and the use of scores for the diagnosis of appendicitis in children”, *European Journal of Pediatric Surgery*, vol. 22, no. 03, 195, 2012. DOI: `10.1055/s-0032-1320017`.

[337] I. Gendel, M. Gutermacher, G. Buklan, L. Lazar, D. Kidron, H. Paran, and I. Erez, “Relative value of clinical, laboratory and imaging tools in diagnosing pediatric acute appendicitis”, *European Journal of Pediatric Surgery*, vol. 21, no. 4, 229, 2011. DOI: `10.1055/s-0031-1273702`.

[338] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention”, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, 2018. DOI: `10.1109/cvpr.2018.00577`.

[339] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 770. DOI: `10.1109/CVPR.2016.90`.

[340] M. Kryzauskas, D. Danys, T. Poskus, S. Mikalauskas, E. Poskus, V. Jotautas, V. Beisa, and K. Strupas, “Is acute appendicitis still misdiagnosed?”, *Open Medicine*, vol. 11, no. 1, 231, 2016. DOI: `10.1515/med-2016-0045`.

[341] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency”, *The Annals of Statistics*, vol. 29, no. 4, 1165, 2001.

[342] Y. Sawada and K. Nakamura, “Concept bottleneck model with additional unsupervised concepts”, *IEEE Access*, vol. 10, 41758, 2022. DOI: `10.1109/ACCESS.2022.3167702`.

[343] M. Yüksekgönül, M. Wang, and J. Zou, “Post-hoc concept bottleneck models”, in *The 11th International Conference on Learning Representations, ICLR 2023*, OpenReview.net, 2023.

[344] E. Marconato, A. Passerini, and S. Teso, “GlanceNets: Interpretable, leak-proof concept-based models”, in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, 21212.

[345] M. Havasi, S. Parbhoo, and F. Doshi-Velez, “Addressing leakage in concept bottleneck models”, in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, 23386.

[346] T. Reddan, J. Corness, K. Mengersen, and F. Harden, “Ultrasound of paediatric appendicitis and its secondary sonographic signs: Providing a more meaningful finding”, *Journal of Medical Radiation Sciences*, vol. 63, no. 1, 59, 2016. DOI: `10.1002/jmrs.154`.

[347] Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks”, in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017, 4885.

[348] S. Shin, Y. Jo, S. Ahn, and N. Lee, "A closer look at the intervention procedure of concept bottleneck models", in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202, PMLR, 2023, 31504.

[349] M. Monteiro, L. Le Folgoc, D. Coelho de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker, "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, 12756.

[350] E. Kim, D. Jung, S. Park, S. Kim, and S. Yoon, "Probabilistic concept bottleneck models", in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202, PMLR, 2023, 16521.

[351] T. P. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models", in *The 11th International Conference on Learning Representations, ICLR 2023*, OpenReview.net, 2023.

[352] R. Marcinkevičs, S. Laguna, M. Vandenhirtz, and J. E. Vogt, "Beyond concept bottleneck models: How to make black boxes intervenable?", in *NeurIPS 2023 Workshop on XAI in Action: Past, Present, and Future Applications*, 2023. DOI: `10.48550/arXiv.2401.13544`.

[353] A. Abid, M. Yuksekgonul, and J. Zou, "Meaningfully debugging model mistakes using conceptual counterfactual explanations", in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, 66.

[354] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision", in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, 8748.

[355] I. Sheth, A. A. Rahman, L. R. Sevyeri, M. Havaei, and S. E. Kahou, "Learning from uncertain concepts via test time interventions", in *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[356] K. Chauhan, R. Tiwari, J. Freyberg, P. Shenoy, and K. Dvijotham, "Interactive concept bottleneck models", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 5, 5948, 2023. DOI: `10.1609/aaai.v37i5.25736`.

[357] D. Steinmann, W. Stammer, F. Friedrich, and K. Kersting, *Learning to intervene on concept bottlenecks*, arXiv:2308.13453, 2023. DOI: `10.48550/arXiv.2308.13453`.

[358] M. Moradi Fard, T. Thonet, and E. Gaussier, "Deep k-means: Jointly clustering with k-means and learning representations", *Pattern Recognition Letters*, vol. 138, 185, 2020. DOI: `10.1016/j.patrec.2020.07.028`.

[359] M. Jia, B.-C. Chen, Z. Wu, C. Cardie, S. Belongie, and S.-N. Lim, *Rethinking nearest neighbors for visual classification*, arXiv:2112.08459, 2021. DOI: `10.48550/arXiv.2112.08459`.

[360] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison”, in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI’19/IAAI’19/EAAI’19, Honolulu, Hawaii, USA: AAAI Press, 2019. DOI: `10.1609/aaai.v33i01.3301590`.

[361] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports”, *Scientific Data*, vol. 6, no. 1, 2019. DOI: `10.1038/s41597-019-0322-0`.

[362] S. Ruder, *An overview of multi-task learning in deep neural networks*, arXiv:1706.05098, 2017. DOI: `10.48550/arXiv.1706.05098`.

[363] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang, “Editing large language models: Problems, methods, and opportunities”, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, 2023, 10222. DOI: `10.18653/v1/2023.emnlp-main.632`.

[364] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2019, `http://www.fairmlbook.org`.

[365] M. Kearns, “Fair algorithms for machine learning”, in *Proceedings of the 2017 ACM Conference on Economics and Computation*, ser. EC ’17, Cambridge, Massachusetts, USA: Association for Computing Machinery, 2017, 1. DOI: `10.1145/3033274.3084096`.

[366] D. S. Char, N. H. Shah, and D. Magnus, “Implementing machine learning in health care—addressing ethical challenges”, *New England Journal of Medicine*, vol. 378, no. 11, 981, 2018. DOI: `10.1056/nejmp1714229`.

[367] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaney-Israni, and A. Goldenberg, “Do no harm: A roadmap for responsible machine learning for health care”, *Nature Medicine*, vol. 25, no. 9, 1337, 2019. DOI: `10.1038/s41591-019-0548-6`.

[368] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations”, *Science*, vol. 366, no. 6464, 447, 2019. DOI: `10.1126/science.aax2342`.

[369] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, “Ensuring fairness in machine learning to advance health equity”, *Annals of Internal Medicine*, vol. 169, no. 12, 866, 2018. DOI: `10.7326/m18-1990`.

[370] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis”, *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, 12592, 2020. DOI: `10.1073/pnas.1919012117`.

[371] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, “CheXclusion: Fairness gaps in deep chest X-ray classifiers”, in *Biocomputing 2021: Proceedings of the Pacific Symposium*, 232. DOI: `10.1142/9789811232701_0022`.

[372] L. Seyyed-Kalantari, H. Zhang, M. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations", *Nature Medicine*, vol. 27, no. 12, 2176, 2021. DOI: 10.1038/s41591-021-01595-0.

[373] R. Marcinkevičs, E. Ozkan, and J. E. Vogt, "Debiasing deep chest X-ray classifiers using intra- and post-processing methods", in *Proceedings of the 7th Machine Learning for Healthcare Conference*, Z. Lipton, R. Ranganath, M. Sendak, M. Sjoding, and S. Yeung, Eds., ser. Proceedings of Machine Learning Research, vol. 182, 2022, 504. DOI: 10.48550/arXiv.2208.00781.

[374] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The measure and mismeasure of fairness", *Journal of Machine Learning Research*, vol. 24, no. 312, 1, 2023.

[375] Y. Savani, C. White, and N. S. Govindarajulu, "Intra-processing methods for debiasing neural networks", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, 2798.

[376] P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser, "A survey of bias in machine learning through the prism of statistical parity", *The American Statistician*, vol. 76, no. 2, 188, 2022. DOI: 10.1080/00031305.2021.1952897.

[377] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning", in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[378] C. Reimers, P. Bodesheim, J. Runge, and J. Denzler, "Conditional adversarial debiasing: Towards learning unbiased classifiers from biased data", in *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021*, Bonn, Germany: Springer-Verlag, 2021, 48. DOI: 10.1007/978-3-030-92659-5_4.

[379] M. B. Zafar, I. Valera, M. Gomez-Rogriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification", in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, PMLR, 2017, 962.

[380] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification", *Journal of Machine Learning Research*, vol. 20, no. 75, 1, 2019.

[381] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification", in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19, Honolulu, HI, USA: Association for Computing Machinery, 2019, 247. DOI: 10.1145/3306618.3314287.

[382] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*, arXiv:1810.01943, 2018. DOI: 10.48550/arXiv.1810.01943.

[383] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems*, vol. 33, no. 1, 1, 2011. DOI: 10.1007/s10115-011-0463-8.

[384] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations", in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., ser. Proceedings of Machine Learning Research, vol. 28, Atlanta, Georgia, USA: PMLR, 2013, 325.

[385] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[386] L. E. Celis, V. Keswani, and N. Vishnoi, "Data preprocessing to mitigate bias: A maximum entropy based approach", in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, 1349.

[387] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer", in. Springer Berlin Heidelberg, 2012, 35. DOI: `10.1007/978-3-642-33486-3_3`.

[388] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning", in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18, New Orleans, LA, USA: Association for Computing Machinery, 2018, 335. DOI: `10.1145/3278721.3278779`.

[389] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification", in *2012 IEEE 12th International Conference on Data Mining*, IEEE, 2012, 924. DOI: `10.1109/ICDM.2012.45`.

[390] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[391] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[392] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction", *npj Digital Medicine*, vol. 4, no. 1, 2021. DOI: `10.1038/s41746-021-00455-y`.

[393] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence", *Nature*, vol. 616, no. 7956, 259, 2023. DOI: `10.1038/s41586-023-05881-4`.

[394] I. Allaouzi and M. Ben Ahmed, "A novel approach for multi-label chest X-ray classification of common thorax diseases", *IEEE Access*, vol. 7, 64279, 2019. DOI: `10.1109/ACCESS.2019.2916849`.

[395] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction", in *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal, Eds., ser. Proceedings of Machine Learning Research, vol. 121, PMLR, 2020, 136.

[396] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs", *Scientific Reports*, vol. 10, no. 1, 2020. DOI: `10.1038/s41598-020-70479-z`.

[397] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*, arXiv:1711.05225, 2017. DOI: `10.48550/arXiv.1711.05225`.

[398] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases", in *Advances in Computer Vision and Pattern Recognition*. Springer International Publishing, 2019, 369. DOI: `10.1007/978-3-030-13969-8_18`.

[399] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study", *PLOS Medicine*, vol. 15, no. 11, 1, 2018. DOI: `10.1371/journal.pmed.1002683`.

[400] H. Zhang, N. Dullerud, K. Roth, L. Oakden-Rayner, S. Pfohl, and M. Ghassemi, "Improving the fairness of chest X-ray classifiers", in *Proceedings of the Conference on Health, Inference, and Learning*, G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, Eds., ser. Proceedings of Machine Learning Research, vol. 174, PMLR, 2022, 204.

[401] Y. Zong, Y. Yang, and T. M. Hospedales, "MEDFAIR: benchmarking fairness for medical imaging", in *11th International Conference on Learning Representations, ICLR 2023*, OpenReview.net, 2023.

[402] F. Esposito, D. Malerba, G. Semeraro, and J. Kay, "A comparative analysis of methods for pruning decision trees", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, 476, 1997. DOI: `10.1109/34.589207`.

[403] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, *A survey of model compression and acceleration for deep neural networks*, arXiv:1710.09282, 2017. DOI: `10.48550/arXiv.1710.09282`.

[404] D. Blalock, J. J. Gonzalez Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?", in *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., vol. 2, 2020, 129.

[405] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage", in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2, Morgan-Kaufmann, 1989.

[406] B. Hassibi and D. Stork, "Second order derivatives for network pruning: Optimal brain surgeon", in *Advances in Neural Information Processing Systems*, S. Hanson, J. Cowan, and C. Giles, Eds., vol. 5, Morgan-Kaufmann, 1992.

[407] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks", in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[408] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference", in *5th International Conference on Learning Representations, ICLR 2017*, OpenReview.net, 2017.

[409] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks", in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, 1398. DOI: 10.1109/ICCV.2017.155.

[410] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network", *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, 30071, 2020. DOI: 10.1073/pnas.1907375117.

[411] O. Antverg and Y. Belinkov, "On the pitfalls of analyzing individual neurons in language models", in *10th International Conference on Learning Representations, ICLR 2022*, OpenReview.net, 2022.

[412] K. Leino, S. Sen, A. Datta, M. Fredrikson, and L. Li, "Influence-directed explanations for deep convolutional networks", in *2018 IEEE International Test Conference (ITC)*, 2018, 1. DOI: 10.1109/TEST.2018.8624792.

[413] K. Dhamdhere, M. Sundararajan, and Q. Yan, "How important is a neuron?", in *7th International Conference on Learning Representations, ICLR 2019*, OpenReview.net, 2019.

[414] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[415] W.-J. Nam, S. Gur, J. Choi, L. Wolf, and S.-W. Lee, "Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 3, 2501, 2020. DOI: 10.1609/aaai.v34i03.5632.

[416] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, vol. 15, no. 56, 1929, 2014.

[417] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning", *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 3, e1452, 2022. DOI: 10.1002/widm.1452.

[418] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid", in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, 202.

[419] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing", *Decision Support Systems*, vol. 62, 22, 2014. DOI: 10.1016/j.dss.2014.03.001.

[420] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database", *Scientific Data*, vol. 3, no. 1, 2016. DOI: 10.1038/sdata.2016.35.

[421] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets", *Journal of Biomedical Informatics*, vol. 83, 112, 2018. DOI: 10.1016/j.jbi.2018.04.007.

[422] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset", *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-11012-2.

[423] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in *3rd International Conference on Learning Representations, ICLR 2015*, OpenReview.net, 2015.

[424] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data", in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE, 2019, 9004. DOI: `10.1109/CVPR.2019.00922`.

[425] J. Lee, E. Kim, J. Lee, J. Lee, and J. Choo, "Learning debiased representation via disentangled feature augmentation", in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, 25123.

[426] M. Vandenhirtz, L. Manduchi, R. Marcinkevičs, and J. E. Vogt, *Signal is harder to learn than bias: Debiasing with focal loss*, arXiv:2305.19671, 2023. DOI: `10.48550/arXiv.2305.19671`.

[427] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*, arXiv:1602.07360, 2016. DOI: `10.48550/arXiv.1602.07360`.

[428] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 2261. DOI: `10.1109/CVPR.2017.243`.

[429] M. Paganini, *Prune responsibly*, arXiv:2009.09936, 2020. DOI: `10.48550/arXiv.2009.09936`.

[430] Z. Liu, E. Gan, and M. Tegmark, "Seeing is believing: Brain-inspired modular training for mechanistic interpretability", *Entropy*, vol. 26, no. 1, 2024. DOI: `10.3390/e26010041`.

[431] A. Madsen, H. Lakkaraju, S. Reddy, and S. Chandar, *Interpretability needs a new paradigm*, arXiv:2405.05386, 2024. DOI: `10.48550/arXiv.2405.05386`.

[432] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, *Generating counterfactual explanations with natural language*, arXiv:1806.09809, 2018. DOI: `10.48550/arXiv.1806.09809`.

[433] L. Bereska and E. Gavves, *Mechanistic interpretability for AI safety – a review*, arXiv:2404.14082, 2024. DOI: `10.48550/arXiv.2404.14082`.

[434] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, "A survey of uncertainty in deep neural networks", *Artificial Intelligence Review*, vol. 56, no. S1, 1513, 2023. DOI: `10.1007/s10462-023-10562-9`.

[435] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, 4396, 2023. DOI: `10.1109/TPAMI.2022.3195549`.

[436] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ICA: A unifying framework", in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, S. Chiappa and R. Calandra, Eds., ser. Proceedings of Machine Learning Research, vol. 108, PMLR, 2020, 2207.

[437] K. Amara, R. Ying, Z. Zhang, Z. Han, Y. Shan, U. Brandes, S. Schemm, and C. Zhang, *GraphFramEx: Towards systematic evaluation of explainability methods for graph neural networks*, arXiv:2206.09677, 2022. DOI: `10.48550/arXiv.2206.09677`.

[438] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, and L. Sigal, “Interpretable spatio-temporal attention for video action recognition”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

CURRICULUM VITAE

PERSONAL DATA

Name	Ričards Marcinkevičs
Date of Birth	December 28, 1995
Place of Birth	Rīga, Latvia
Citizen of	Latvia

EDUCATION

2017 – 2019	ETH Zurich, Zürich, Switzerland *Final degree:* Master of Science ETH in Statistics
2014 – 2017	Maastricht University, Maastricht, Netherlands *Final degree:* Bachelor of Science in Data Science and Knowledge Engineering
2009 – 2014	Rīga Secondary School 34, Rīga, Latvia *Final degree:* General Certificate of Secondary Education
2009 – 2014	Rīga Secondary School 95, Rīga, Latvia

EMPLOYMENT

2019 –	Doctoral Researcher *ETH Zurich*, Zürich, Switzerland
2015 – 2017	Research Intern *Medtronic*, Maastricht, Netherlands

Framing the question as the choice between accuracy and interpretability is an incorrect interpretation of what the goal of a statistical analysis is. The point of a model is to get useful information about the relation between the response and predictor variables. Interpretability is a way of getting information.

Leo Breiman [10]