Learning Rule-based Decision-Making Systems from Heterogeneous Longitudinal Data

Ameer Hamza Shakur

A dissertation proposal
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Shuai Huang, Chair

Archis Ghate

Chaoyue Zhao

Cynthia Chen

Peter Tarczy-Hornoch

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Learning Rule-based Decision-Making Systems from Heterogeneous Longitudinal Data

Ameer Hamza Shakur

Chair of the Supervisory Committee:
Shuai Huang
Industrial and Systems Engineering

Recent advances in sensing technology have greatly expanded our capacities to collect data from a diverse pool of patients in unprecedented spatial-temporal resolutions and from a variety of different sources. These technological advances enormously increase the complexity of modern patient data sets and have thrown up many challenges and opportunities for analysis, as the methodological framework of the classic models designed to model the average effects are found to be over-simplified. The increasing size and dimensionality of modern data sets also make the development of sparse models imperative. There has been a recent push towards interpretable models that can not only provide accurate predictions but also explain why the prediction has been made. Interpretability and explainability hold the key to their success in complex applications such as healthcare so that the model decisions may be communicated to, and evaluated by medical professionals, and enhance accountability. A faithful understanding of the uncertainty in the predictions is becoming critical as decision-making can be dangerous and expensive in such applications especially as more and more systems are getting automated in this age of data. Larger datasets often bring increasing heterogeneity of data, so personalized decision models that can recognize heterogeneity between observations and subgroups are important in medical applications. Additionally, complex data structures such as survival data often have incomplete data that must be carefully modeled and poses their own challenges. Further, these data are often multimodal and may come in various

representations such as text, audio-visual, or time series.

This dissertation focuses on developing rules-based interpretable machine learning models that can address these new and exciting challenges in modern datasets. First, we introduce SURVFIT, a "doubly sparse" rule extraction formulation for survival data. This doubly sparse method can induce sparsity both in the number of rules and in the number of variables involved in the rules. Further, a systematic rule evaluation framework that includes statistical testing, decomposition analysis, and sensitivity analysis is developed to assist the interpretability of the extracted rules. Our next contribution, GPSRL, proposes a Bayesian semi-parametric ordered rule-list methodology to address the heterogeneity and quantify uncertainty. The use of ordered rule lists enables us to model the heterogeneity while keeping in check the model complexity. We apply these methodologies to real-world applications through a sepsis survival dataset. Finally, we explore the applications of the rules-based approach to the discovery of multimodal biomarkers in ADHD. We identify interesting interactions among two modalities of data — eye movement patterns and EEG signals. The detection of these interactions would help us better understand the condition and develop better prediction models and intervention strategies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my advisor, Professor Shuai Huang for his guidance and continuous encouragement over the course of my PhD years. Shuai's generosity with his time and his trust has been crucial to my progress as I undertook this challenging task. I will always be indebted to him for his influence on my development, both as a professional and as an individual.

A special thank you to my committee members, Professor Cynthia Chen, Professor Archis Ghate, Professor Peter-Tarczy Hornoch, and Professor Chaoyue Zhao. Thank you for your insightful comments, and your support throughout this journey.

I also owe thanks to all my collaborators over the years, Dr. Xiangyu Chang, Dr. Xiaoning Qian, Dr. Cynthia Chen, Dr. Ji-Eun Kim, Tianchen Sun, and others. None of this would have been possible without your many invaluable contributions to this work.

To the staff at the ISE department, Neelu Rajvanshi, Sheila Prusa, and Jennifer Tsai, I am grateful for all your support throughout my time here.

To Dr. Victoria Diaz, Dr. Jingshuo Feng, Dr. Fiete Krutein, Serin Lee, Bandhav Veluri, and all my friends and colleagues at ISE, past and present — I will deeply cherish and remember both the long discussions we have had, and the laughs we have shared.

To my parents, Jafferunnisa Begum and Abdul Shukur, I owe you for everything that I am. I also need to mention my three wonderful siblings, Fatima-tuz-Zahra, Abdul Wahid and Abdul Hameed, thank you for having my back no matter what.

DEDICATION

To my father, for setting an example worth emulating

Chapter 1

INTRODUCTION

1.1 Motivation

Far-reaching advances in sensing technology and experimental instruments have greatly expanded our capacities to collect data from a diverse pool of patients, as well as from a large variety of sensors and sources. These technological advances also enormously increase the complexity of modern patient medical data and have thrown up many challenges and opportunities for analysis. These challenges may come in various forms: firstly, modern datasets are high-dimensional consisting of a large number of variables - which motivates using *sparse* models that can identify the relevant variables to be used, instead of using all the variables which may lead to the issue of *overfitting* [1]. Secondly, as the machine learning community has also realized, *interpretable* models that can not only provide accurate predictions but also explain why the prediction has been made hold the key for their success in complex applications such as healthcare so that the model decisions may be communicated to, and evaluated by medical professionals. Due to the nature of many biological mechanisms, models that are able to characterize some relationship between the statistical factors have shown remarkable reproducibility in genomics studies [2] and have led to the identification of highly reproducible biomarkers that are strong predictors of breast cancer [3]. Thirdly, a faithful understanding of the *uncertainty* in the predictions of survival rate is critical in many applications - both medical and engineering, as the interventions staged to manage failure might be dangerous, expensive, or both. Therefore, developing models that can account for the uncertainty in their predictions is highly desirable. Indeed, accounting for model uncertainty has also been found to enhance predictive performance [4]. Fourthly, many of these datasets are *multimodal* in nature and consist of information that could be in different representations

such as text, audio-visual information, various physiological measurements, etc. Integrating these modalities in a coherent manner would lead to better model performance, as the information from different modalities may augment one another. This would also improve the robustness of the models since weak or noisy signals from one modality might be overcome by information coming from other modalities. Perhaps even more interestingly, multimodal data analysis also has the potential to discover unknown characteristics or relationships, and drive clinical and biological discovery. Furthermore, there are various challenges associated with the complexity of the data itself. For example, data in survival analysis often has censored observations with incomplete information. In survival analysis, the goal is to model the *failure time* of an event — the event usually being morbidity or mortality in medical applications or break-down of some equipment in engineering. It has been used to study how statistical factors influence the mortality for a diverse range of diseases, e.g. congestive heart failure [5], pediatric trauma [6], lupus [7], breast cancer [8], to name but a few. Risk assessment is also helpful in the early detection and diagnosis of complex diseases and helpful for doctors to monitor the disease progression [9]. On the engineering side, survival analysis has been used for applications such as water network assessment to analyze deterioration in the network to proactively assess the risk of failure [10], studying mechanical reliability of new generation ceramics [11], customer retention [12], equipment failure [13] etc. The challenges associated with survival analysis, e.g., when applied to medical data, may stem from the complexity of the underlying processes, the difficulty of data collection, and the incompleteness of time-to-event information. This incompleteness, called *censoring* can be a result of either the subject leaving the study mid-way or the study being terminated before an event is experienced by the subject. Therefore, it is also important to develop models that best use incomplete information and such data structures. This dissertation aims to develop a rules-based framework to address some of these challenges presented by modern patient datasets, with a particular focus on survival analysis and the discovery of multimodal biomarkers.

1.2 Literature review

This dissertation focuses on the challenges and opportunities promised by modern patient medical datasets. Here we present some of the current literature in this domain to place our work in context.

1.2.1 Interpretable and Rule Based Modeling

Rule-learning methods have been one main interest of artificial intelligence since their inception, using heuristic algorithms [14, 15], or logic deduction approaches [16, 17] to learn a set of predictive rules from data. These early efforts had been largely limited by the daunting computational demand of learning rules from data, i.e., it is essentially a combinatorial optimization problem, not to mention the equally challenging aspect of developing a sound and statistically solid formulation to guide the learning. In recent years, rule-based modeling has seen a renewed resurgence with the growing interest in interpretability, and the development of tree ensembles such as random forests [18] and sparse regularization techniques such as LASSO [19]. The rulefit model [20] cleverly combines these methods by first generating a huge list of rules from a tree ensemble, and then applying LASSO to select a minimum set of rules that can predict the outcome with a good accuracy. Building on these new developments, rule-based models have served to understand risk-predictive profile patterns and build predictive models for diseases, such as Type 1 diabetes [21], Type 2 diabetes [22], depression [21], classification of cancer gene expression data [23], sepsis [24] etc. However, there are only a few works [25, 26] that propose rule learning for survival analysis. In [25], the idea of rulefit [20] is extended to survival analysis using the Cox regression loss function with the lasso regularization, while [26] present a heuristic rule induction approach based on a separate-and-conquer strategy in combination with the log-rank [27, 28] statistical test. Surprisingly, sparsity of variables involved in the rules has not yet been addressed in rule-learning literature. In Chapter 2 we propose a "doubly sparse" rule extraction formulation for survival data to address this particular gap. This doubly sparse method can induce

sparsity both in the number of rules and in the number of variables involved in the rules. Our method has the computational efficiency needed to realistically solve the problem of rule extraction from survival data if we consider both rule sparsity and variable sparsity, by adopting a quadratic loss function with an overlapping group regularization. Further, we also present a systematic rule evaluation framework that includes statistical testing, decomposition analysis, and sensitivity analysis of the factors involved in the rules to understand secondary effects.

1.2.2 Modeling Uncertainty and Heterogeneity in Patient Data

Standard survival models assume independent and identically distributed data observations, however, this is rarely the case in real-life medical data, different individuals may react very differently to the same condition or illness and often of different risk profiles. This is a source of heterogeneity, particularly in large datasets, and modeling of heterogeneous effects has been found to have a significant impact on survival and hazard estimates [29]. Unobserved heterogeneous effects have been popularly modeled as *frailty* models [30], where an individual's frailty are the random effects that act on their hazard functions. While frailty models are able to tackle unobservable heterogeneity, occurring due to the different risk attributes of each individual, another major source of heterogeneity is due to observations belonging to different populations where heterogeneity cannot be fully captured by random effects. In this scenario, latent class mixture modeling [31] approaches offer a way of incorporating additional heterogeneity and partition the dataset to discover sub-groups of the population that are homogeneous and share the same risk profiles. Under this paradigm, rule-based learning uses spatial partitioning to model heterogeneous data. Rule-based methods have been shown to be effective in identifying subgroups with heterogeneous risk profiles in a patient population [32]. Greedy decision tree models such as CART [33] are typical examples, but they provide a quite restrictive result, i.e., the partitioning is sub-optimal. Optimal partitioning such as through integer programming [34] or Bayesian decision trees [35] is NP-hard and computationally demanding due to the exponential search space which

severely limits both the depth of the tree and the number of variables that can be considered. Comparing with the tree models, the rule-based methods, which work on the same principle of partitioning, have found larger flexibility and efficiency in a range of applications. As the purpose of using partitioning is to tackle patient heterogeneity, it can be used to group data into subsets with similar response characteristics that enable the sub-grouping of subjects and subsequent separate modeling of each subgroup. Rules can also be flexibly used or re-organized to make better decisions, one example is the recent development of Bayesian rule lists (BRL) for classification [36] that is able to incorporate Bayesian modeling to quantify uncertainty. BRL is able to strike a balance between greedy and optimal partitioning and provides good generalizability with significantly lower computational load. In Chapter 3 we propose an integrative framework that uses ordered rule lists to derive a rule-based decision-making approach, while within the regime defined by each rule, survival risk is modeled via a Gaussian process latent variable model. The computational challenges are overcome by a tailored Markov Chain Monte Carlo algorithm with a nested Laplace approximation for the latent variable model to search over the posterior of the rule lists efficiently.

1.2.3 Discovery of multimodal biomarkers

Due to the complexity of biological processes, it is reasonable to think that one modality might not be sufficient in explaining the process entirely. Relying on a combination of modalities may provide a more comprehensive view of the underlying process and hence lead to better and more effective decision-making outcomes. Patient data can come in several representations such as text in the form of medical reports or prescriptions, images in the form of MRI and brain scans, signals from various different types of sensors such as EEG, or eye-movement data, or longitudinal physiological measurements such as heart rate and blood pressure, or genetic repositories etc. Using multiple modalities in machine learning has been quite effective in various applications such as improving cardiovascular disease care [37], prediction of symptoms in Parkinson's' patients [38], joint diagnosis of human cancers [39] and advancing precision health [40] and oncology [41] to name just a few. Multimodal

models have recently found success in the study of neurological disorders, as the fusion of fMRI data and grey matter volume data has enabled the discrimination between autism spectrum disorder and schizophrenia to a much greater extent as compared to using just a single modality [42]. There is a large body of work in the research literature on the use of electroencephalograph (EEG) for the diagnosis of attention deficit hyperactive disorder (ADHD) [43, 44, 45, 46]. In recent times, there have also been several works developing ADHD detection systems using eye-movement patterns. [47, 48, 49]. However, these two important modalities have so far not been used together for the detection of ADHD or biomarkers of ADHD, though the multimodal literature focused on the fusion of fMRI and EEG data [50, 43]. There is some prior work [51] to show that characteristics of eye movements and EEG are complementary to the task of emotion recognition and data fusion could significantly improve emotion recognition accuracy in comparison with single modality. This suggests that these two modalities may also be important in the diagnosis of ADHD as it is strongly correlated with emotional dysregulation and the intensity of emotions [52]. In Chapter 4, we present a rule-based approach to detecting multimodal biomarkers of ADHD using both these modalities. We simultaneously collected both eye-movement patterns and EEG signal data while the subject is carrying out a specific task. Then, we extracted the most relevant features from each of these modes of data, and demonstrate how rule-learning can be used to detect significant multimodal biomarkers. To the best of our knowledge, our work is the first attempt to explore and identify interesting interactions among two modalities of data, eye movement data, and the EEG signal, for biomarker discovery for ADHD.

1.3 Research Objectives

The objectives of this research proposal are:

- Develop interpretable and novel computational models and algorithms for survival modeling with a focus on complex challenges such as interpretabilty and heterogeneity

- Apply the proposed statistical models for knowledge discovery on real-life health datasets.

- Identify novel biomarkers and interactions of ADHD using a multimodal dataset containing eye-movement patterns and EEG signals

1.4 Organization of the dissertation

This dissertation is an attempt to address several of the aforementioned challenges in knowledge extraction and modeling with modern patient datasets through novel modeling techniques. Chapter 2 provides our first contribution, an interpretable rule extraction approach that induces variable sparsity among the rules. Next, Chapter 3 presents a semi-parametric Bayesian survival rule list framework that is able to model heterogeneity and uncertainty. In Chapter 4, we demonstrate the utility of the rules-based approach to discover multimodal biomarkers of ADHD using EEG and eye-movement data. Finally, chapter 5 concludes this dissertation and points towards future areas of research.

Chapter 2

DOUBLY SPARSE RULE LEARNING FOR SURVIVAL DATA

Survival data analysis has been leveraged in medical research to study disease morbidity and mortality, and to discover significant bio-markers affecting them. A crucial objective in studying high-dimensional medical data is the development of inherently interpretable models that can efficiently capture sparse underlying signals while retaining high predictive accuracy. Recently developed rule ensemble models have been shown to effectively accomplish this objective; however, they are computationally expensive when applied to survival data and do not account for sparsity in the number of variables included in the generated rules. To address these gaps, we present SURVFIT, a "doubly sparse" rule extraction formulation for survival data. This doubly sparse method can induce sparsity both in the number of rules and in the number of variables involved in the rules. The computational superiority of our model is a result of our computational formulation that adopts a quadratic loss function with an overlapping group regularization. Further, a systematic rule evaluation framework that includes statistical testing, decomposition analysis, and sensitivity analysis is provided. We demonstrate the utility of SURVFIT via experiments carried out on a synthetic dataset and a sepsis survival dataset from MIMIC-III.

2.1 Introduction

When analyzing biological and medical datasets, an often encountered scenario is the need to simultaneously analyze multiple variables and understand their impact on a certain disease or biological condition. In this endeavor, regression methods have been a typical approach. These methods help us understand the relative importance of variables primarily in terms of their average effects on the outcome rather than their synergistic interactions. Though

adding interaction terms to the regression model can certainly enable their application in evaluating the significance of these interaction terms, regression models themselves are not adequate for discovering such interactions due to both computational and statistical challenges, i.e., the number of potential interactions grows at a super-exponential rate regarding the number of variables. The rule learning approach is a natural way to address these challenges. An old song since its inception in the early 70s and 80s as a typical approach of Artificial Intelligence, it now finds its new tune in the 21st century as a result of considerable developments in the fields of machine learning and optimization such as random forests [18] and sparse regularization models such as LASSO [19]. Rulefit [20] is a good example of a model that cleverly combines these methods by first generating a huge list of rules from a tree ensemble, and then applying LASSO to select a minimum set of rules that can predict the outcome with good accuracy. Compared to rule learning methods developed before Rulefit that mostly used heuristic algorithms [14, 15] or logic deduction approaches [16, 17] to derive rules, Rulefit is both computationally efficient, inherited from random forest and LASSO, and statistically well justified, as random forest uses bootstrap aggregation to generate an ensemble of tree models and has the ability to cover a wider range of the rule space, therefore being less susceptible to being stuck in local optima. An additional advantage of applying rule-based models to biomedicine is that they can be easily communicated to, and evaluated by medical professionals. Several recent works have successfully applied rule-based models to diverse biomedical datasets to understand risk-predictive profile patterns and build predictive models for diseases, including Type 1 diabetes [21, 32], Type 2 diabetes [22], depression [53], classification of cancer gene expression data [23] and detection of predictive rules for Alzheimer's disease [54]. However, these works were not focused on survival data. Survival rule models proposed in literature [25, 26] lack the methodology to impose sparsity on the variables that constitute the rules. Sparsity in variables has been proven to be a main concern in a wide range of applications. Surprisingly, sparsity of variables involved in the rules has not yet been addressed in rule-learning literature. Therefore, our research seeks to address these gaps and focuses on a rule-learning approach that can efficiently learn a

"doubly sparse" set of rules and analyze their properties for survival analysis, a field with critical applications in biomedicine.

In this work, we propose a new rule learning method, SURVFIT, with three main contributions. First, we aim to fill in a gap that concerns rule learning with variable sparsity, i.e. "double sparsity" for survival data analysis. To achieve this, we propose a formulation that adopts a quadratic loss function and an overlapping group regularization term. The quadratic loss function allows us to bypass the partial log-likelihood loss function of the Cox models that has caused considerable computational difficulty for high-dimensional applications, and the proposed regularization enables us to not just select the most important rules but also induce sparsity of variables involved in the selected rules. This "double sparsity", in both rules and variables, has so far not been addressed in the literature of rule learning. Second, we propose and compare different optimization strategies for solving our optimization problem and discuss their advantages and trade-offs. Third, we provide a systematic rule evaluation framework for evaluating and examining the statistical significance of the rules extracted via SURVFIT. The framework includes statistical testing of the rules' ability to discriminate between low-risk and high-risk observations, decomposition analysis, and sensitivity analysis of the cutoff values.

2.2 Background and Limitations

In this section, we introduce certain widely used rule-based analysis methods. Further, we discuss the advantages and disadvantages of each of these methods and proceed to explain their limitations.

2.2.1 Survival Trees.

Parametric (and semi-parametric) regression models impose a specific link function on the response and face challenges in incorporating interactions between variables. Trees provide a flexible approach that can detect interactions in variables without explicitly specifying them beforehand. They also do not assume a specific link function and are widely used as they are

easy to interpret and understand for medical professionals. To build a tree model, the data is split into child nodes in such a way that the subsets of data within each of the child nodes have more homogenous outcome responses, i.e., the outcomes of assigned data samples in a child node would be more similar (similarity is defined by a pre-specified similarity measure). Specifically, a split concerns a variable and a cutoff value, i.e., if x is the selected variable and c is its cutoff value, the splitting decision $x \leq c$ vs. $x > c$ would result in a split of the data into two child nodes. In each split to grow the tree, the choice of variable and cutoff value that makes each child node most homogenous, and each pair of child nodes most dissimilar is sought among all possible choices according to criteria based on the similarity measure. The splitting decisions are recursively applied to each child node to grow the tree until a stopping criterion is obtained. A tree that is grown too deep will overfit to the training data and may be pruned to obtain a smaller tree that generalizes well to new and unseen data. Thus, trees will naturally group together observations with similar outcomes which leads them to be highly interpretable. The main difference between classical decision trees and survival trees is in the splitting criteria used to partition the data - survival trees use criteria that make each child node to be most similar in terms of their survival or hazard functions. Several splitting criteria have been developed for survival trees, e.g., the maximum log-rank statistic [27, 28] and the log-rank score [55].

2.2.2 Random Survival Forests.

Random forests are an ensemble model generated by combining many decision or survival trees where each tree is built on a randomly bootstrapped sample of data and a randomly selected subset of variables. The average outcome of all of these binary trees is the output of the random forest. Random forests were initially developed for regression and classification problems [18] and later extended to apply to survival data, [56, 57] where an ensemble of survival trees is used to build a forest. For a given input, an average of the cumulative hazard prediction of all the survival trees in the ensemble is the output of the random survival forest.

2.2.3 Sparse Regularization.

Modern medical datasets are high-dimensional, which leads to challenges associated with the *curse of dimensionality*, particularly in datasets with many correlated variables. This makes it critical to build models that are sparse with respect to the number of variables and to identify the most significant variables affecting the underlying process. Development of sparse regularization methods for survival analysis is a line of efforts seeking to deal with the challenges of high dimensional data in both regression-based and tree-based methodologies. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ of n observations and m variables where m is large, our goal in variable selection is to choose a number of variables $k(\ll m)$ that are the most significant predictors of the output. Consider a model parametrized by the coefficients corresponding to the variables, $\boldsymbol{\beta}$, and a loss function $\mathcal{L}(\boldsymbol{\beta}; \mathbf{X})$ that is to be minimized to obtain model coefficient estimates. Sparse regularization works by regularizing the loss function with sparsity inducing norms such as the ℓ_1 norm, $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n} |\beta_i|$ that was first proposed as the LASSO model [19] for linear regression. The ℓ_1 norm has the property of shrinking the coefficients closer to zero which enables variable selection by eliminating those variables whose coefficients are nearly zero. Inspired by the success of the sparse regularization approaches, sparse Cox regression models have also been developed [58] to regularize the Cox regression loss. An important direction in sparse regularization is structured sparsity regularization to obtain desired model characteristics such as selection of groups of variables, i.e., selecting all variables in a predefined group of variables or none at all. The Group LASSO [59] solves the group selection problem by using an $\ell_{2,1}$ norm regularization, $\sum_{g=1}^{|G|} \|\boldsymbol{\beta}_g\|_2$, where $\boldsymbol{\beta}_g$ are coefficients in group g, belonging to a set of groups G and the ℓ_2 norm is given by $\|\boldsymbol{\beta}_g\|_2 = \sqrt{\sum_{i \in g} \beta_i^2}$. The Sparse Group LASSO [60] generalizes the Group LASSO to also induce within group sparsity in the solution by using a regularization that is the sum of $\ell_{2,1}$ and ℓ_1 norm, i.e. $\sum_{g=1}^{|G|} \|\boldsymbol{\beta}_g\|_2 + \|\boldsymbol{\beta}\|_1$. As the complexity of the sparsity-inducing norms increases, their adoption to survival regression models such as the Cox regression [61] still poses significant algorithmic challenges despite the computational advantages of

these methods in the typical regression setting. While most of these efforts mainly focus on variable selection when the link function of the model is linear, the advances in sparse regularization approaches also positively impacted the work on tree-based methods in regression and classification like Rulefit [20] as well as in survival analysis, such as the method developed in **pre** [25], where a sparse set of survival rules are generated by constructing an ℓ_1-regularized Cox regression model over an exhaustive set of rules extracted from the data through bootstrapped survival trees. However, the Cox partial likelihood function used in **pre** has difficulty in scaling to high dimensional data. Although the regularized Cox model can handle relatively high dimensions, the optimization algorithms that are built on the Cox partial likelihood function scale poorly when regularized with structured norms such as Group LASSO and Sparse Group LASSO [62].

2.2.4 Our Contributions

We propose a new rule learning method, SURVFIT, with three main contributions. First, we aim to fill in a gap that concerns rule learning with variable sparsity, i.e., "double sparsity" for survival data analysis. To achieve this, we propose a formulation that adopts a quadratic loss function and an overlapping group regularization term. The quadratic loss function allows us to bypass the partial log-likelihood loss function of the Cox models that has caused considerable computational difficulty for high-dimensional applications, and the proposed regularization enables us to not just select the most important rules but also induce sparsity of variables involved in the selected rules. This "double sparsity", in both rules and variables, has so far not been addressed in the literature of rule learning. Second, we propose and compare different optimization strategies for solving our optimization problem and discuss their advantages and trade-offs. Third, we provide a systematic rule evaluation framework for evaluating and examining the statistical significance of the rules extracted via SURVFIT. The framework includes statistical testing of rules' ability to discriminate between low-risk and high-risk observations, decomposition analysis, and sensitivity analysis of the cutoff values. An overall presentation of this framework is shown in Fig. 2.1.

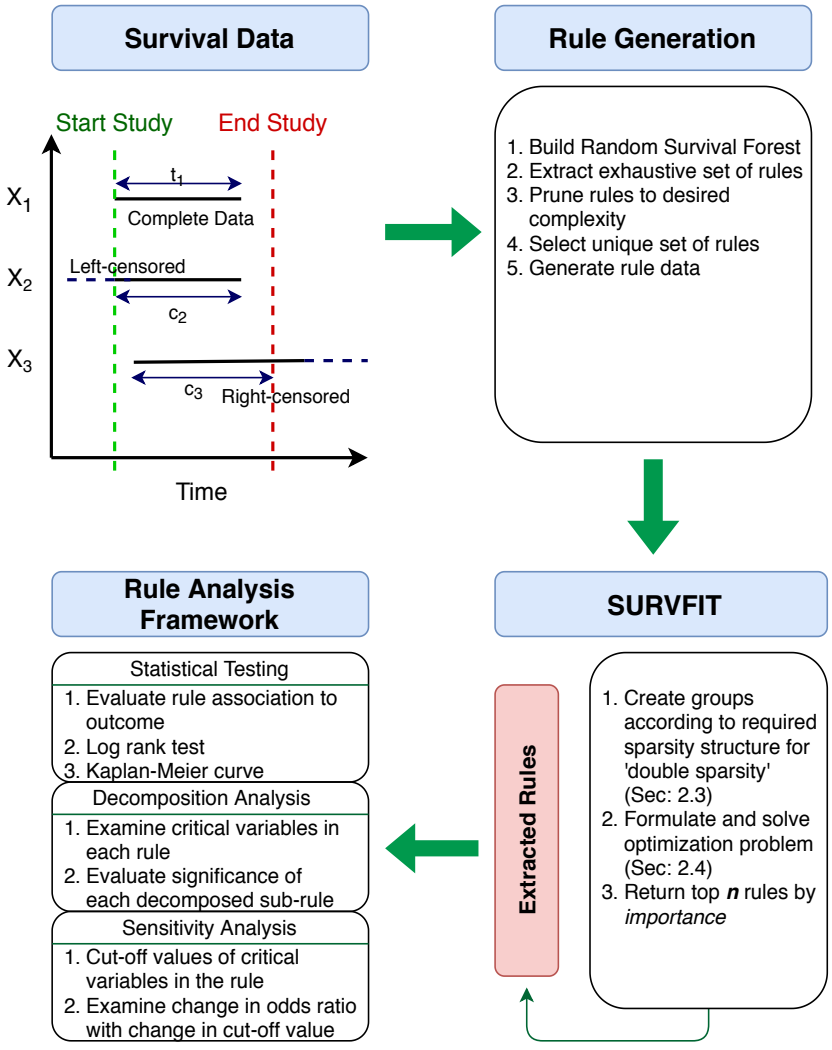


Figure 2.1: A schematic outline of the SURVFIT algorithm

2.3 SURVFIT

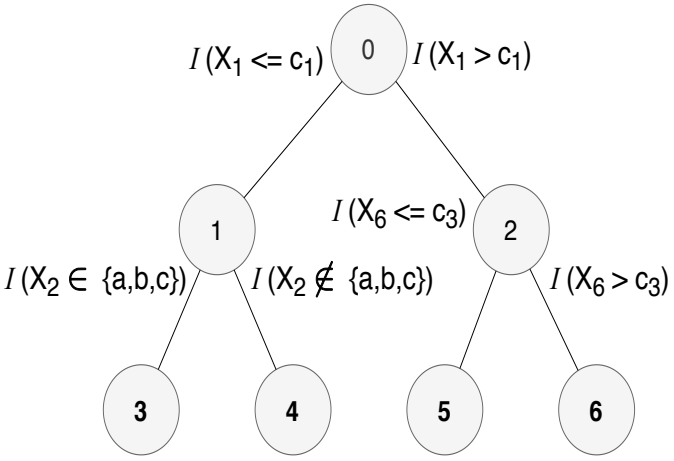
2.3.1 Methodology

Rule learning is a challenging problem mainly due to the combinatorial nature of rules, i.e., a rule is expressed as the product of a few indicator functions $I(\cdot)$ on propositions of values taken by variables in an observation $\mathbf{x}$,

$$r_k(\mathbf{x}) = \prod_{p=1}^{|\mathbf{x}|} I(x_p \in s_{pk}). \tag{2.1}$$

For continuous variables, s_{pk} is a contiguous interval while for categorical variables, it is an explicitly specified set. A rule either gives 0 or 1 as its outcome for an input observation. If its outcome is 1, it means all the conditions on its constituent variables are satisfied, i.e., $\{x_p \in s_{pk}\}_1^{|\mathbf{x}|}$. We say that rule r is endorsed by observation $\mathbf{x}$ if $r(\mathbf{x}) = 1$. Through this combinatorial nature, rules provide effective semantics to capture interactions among variables, not only in the qualitative sense, i.e., which variables interact with which, but also in the quantitative sense, i.e., the cutoff values used in the conditions of the rules. It is also due to this combinatorial nature that rules are information-rich, but computationally and statistically challenging to detect from data. Recent breakthroughs in rule learning benefit from an insight that a decision tree can be readily decomposed into a set of rules as shown in Fig. 2.2. Tree ensemble models such as random forests can therefore be used to generate a huge set of rules. Then, formulations could be developed to filter this set and select a sparse set of the most representative and informative rules. Rule learning methods such as Rulefit [20] and **pre** [25] follow this line. However, these methods do not consider the sparsity in the variables that are involved in the extracted rules. Sparsity of variables has proven to be a critical trait of machine learning models that can achieve robust prediction performance and interpretability in practice. An immediate example that will be shown in the medical application in this paper is that variables collected in healthcare applications are usually highly correlated, and thus it is important to be able to generate rules that involve only a sparse selection of significant variables. For example, two variables may show up in different

rules, though only one variable is truly significant, and the other is redundant.



NODE	EXTRACTED RULES
3	$X_1 <= c_1$ & $X_2 \in \{a,b,c\}$
4	$X_1 <= c_1$ & $X_2 \notin \{a,b,c\}$
5	$X_1 > c_1$ & $X_6 <= c_3$
6	$X_1 > c_1$ & $X_6 > c_3$

Figure 2.2: An example decision tree and corresponding rules extracted from terminal nodes

2.3.2 Rule generation

In order to generate an abundant set of rules that will be pruned by our learning formulation, we use the following algorithm to build the random survival forests.

1. Draw a given number of bootstrap samples from the original data.

2. Grow a survival tree for each bootstrapped sample as follows:

   (a) Use one of the splitting criteria discussed in 2.2.1 to recursively build a tree using a randomly selected subset of variables for each split.

   (b) Grow the tree until no new child nodes can be formed because of the stopping condition that each node must contain a minimum number of unique events.

3. Aggregate all the survival trees to obtain an ensemble.

4. Extract rules of the desired length and complexity from the tree ensemble to generate a large rule list.

2.3.3 The loss function

This initial set of rules is denoted as $\{r_k(\mathbf{x})\}_1^K$, where K is the total number of rules. We then develop a learning formulation to guide the selection of the final rules, which should be a minimum number of rules (i.e., the number should be much smaller than K) that could achieve optimal prediction on the time-to-event outcome of survival data. It is tempting to use the existing Cox proportional hazards regression model and take the K rules as K input variables, then conduct sparse learning on this Cox model-based formulation. This is a reasonable approach, but the partial likelihood function used in the Cox regression model has been found to scale poorly in high-dimensional applications, particularly with complex group norms [62], including the recently developed **pre** [25], a rule learning method for survival analysis, that is also built on the negative partial log-likelihood loss in the Cox proportional hazards model. Thus, we resort to another loss function. In this paper, we are concerned with the linear model, but our method could be extended to nonlinear models as well. The prediction by the linear model is,

$$t(\mathbf{x}_i) = \boldsymbol{\beta}_0 + \sum_{k=1}^{K} \boldsymbol{\beta}_k r_k(\mathbf{x}_i). \tag{2.2}$$

We adopt the robust loss function developed in [63],

$$\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \mathbf{X}) := \sum_{\{i|\delta_i=1\}} \frac{1}{2}\big(t(\mathbf{x}_i) - t_i\big)^2 \tag{2.3}$$

$$+ \sum_{\{i|\delta_i=0\}} \frac{\gamma}{2}\big(\min(0, t(\mathbf{x}_i) - t_i)\big)^2. \tag{2.4}$$

The first term in (2.4) is the least-squared loss that penalizes the difference between the predicted outcome $t(\mathbf{x}_i)$ and the real outcome t_i for each complete observation $\mathbf{x}_i$. The second term is a squared hinge loss which penalizes the predictions for censored data only when the predicted event time $t(\mathbf{x}_j)$ of censored observation $\mathbf{x}_j$ is smaller than the censor time t_j. The penalty is zero when $t(\mathbf{x}_j)$ is greater than t_j. The hyperparameter γ controls the influence of censored data in parameter estimation and is selected via cross-validation.

2.3.4 Doubly Sparse Rule Extraction

Consider the regression optimization problem (2.4) applied to rules, i.e, let $\boldsymbol{\beta}$ be the coefficients of the complete set of rules and $\mathbf{X}$ the rule data matrix. Each column in $\mathbf{X}$ is a binary variable denoting whether or not the observations endorse a rule. To achieve sparsity in both rules and variables, we integrate two regularization terms simultaneously. On one hand, to enforce sparsity on the rules, we adopt the ℓ_1 norm that was used in the least absolute shrinkage and selection operator (LASSO) regularization for regression proposed by [19], i.e., it regularizes the loss function with ℓ_1 norm penalty on the coefficients corresponding to rules ($\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n} |\boldsymbol{\beta}_i|$) to push some coefficients to 0. In the rule learning literature for regression, Rulefit [20] and **pre** [25] utilize LASSO to extract a sparse rule list from a rule ensemble. Following this line, we propose the following formulation to enforce sparsity in the *cardinality* of rules

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ l(\boldsymbol{\beta}, \mathbf{X}) + \lambda \cdot \|\boldsymbol{\beta}\|_1, \tag{2.5}$$

where $l(\boldsymbol{\beta}, \mathbf{X})$ is the squared error loss for regression. As we have mentioned, this formulation does not enforce sparsity among the variables involved in the extracted rules. As one of the main contributions of this work, that is to enforce a "doubly sparse" set of rules, our idea is inspired by the work on sparsity-inducing norms for structured variable selection [64]. A particularly useful concept is the Overlapping Group LASSO regularization. Here, as Fig. 2.3 shows, the rules form a group structure due to their overlapping use of the variables. To exploit this structural property in order to enforce sparsity on the variables, the regularization term we propose to use is, therefore,

$$\Omega(\boldsymbol{\beta}) = \sum_{G \in \mathcal{G}} \|w^G \circ \boldsymbol{\beta}\|_2, \tag{2.6}$$

where $\mathcal{G}$ is a set of overlapping groups of coefficients; $w^G_{G \in \mathcal{G}}$ are $|\boldsymbol{\beta}|$-dimensional vectors such that $w^G_j > 0$ if $j \in G$ and $w^G_j = 0$ otherwise; $\mathbf{x} \circ \mathbf{y}$ denotes element-wise multiplication of two vectors, $\mathbf{x}$ and $\mathbf{y}$. This regularization term $\Omega(\boldsymbol{\beta})$, despite its difference from the ℓ_1 norm from the surface, is like an ℓ_1 norm at the group level to promote group sparsity. A

few different group structures that may be used to obtain specific sparsity patterns were explored by [64], though the challenge of group construction was not discussed since it requires prior information about the problem under consideration and the required sparsity patterns. Since our work has a well-defined goal, i.e., the variable sparsity in a rule ensemble, we can develop a natural way to construct the groups. To do this, we first choose our set of groups $\mathcal{G} = \{G_1, \dots G_P\}$ such that each group G_p, corresponding to the variable x_p, contains the indices of the rules which involve the variable x_p. That is, if there are n_p number of rules containing variable x_p, then $G_p = \{p_1, \dots p_{n_p}\}$, where $\{p_1, \dots p_{n_p}\} \subseteq \{1, \dots K\}$ and all rules r_{p_j} $(1 \leq j \leq n_p)$ contains the variable x_p in at least one of its combinatorial statements. Next, for each group G in $\mathcal{G}$, we define $\boldsymbol{\beta}_G$, a vector in $\mathbb{R}^{|G|}$, that consists of the elements of $\boldsymbol{\beta}$ belonging to G. Now, the formulation (4.2) could be further developed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \mathbf{x}) + \lambda_1 \cdot \|\boldsymbol{\beta}\|_1 + \lambda_2 \cdot \sum_{p=1}^{P} q_p \|\boldsymbol{\beta}_{G_p}\|_2. \tag{2.7}$$

Note that here q_p is an optional weight for each group. In our case, $q_p = \sqrt{1/|G_p|}$ to normalize the penalty term for groups of varying sizes. The hyperparameters λ_1 and λ_2 can be determined by cross-validation. The obtained solution is such that the potential nonzero patterns in the model are a complement of an intersection of a subset of groups. Fig 2.3 provides a representation of the sparsity structure that this regularization will induce. For example, if group 4 is left out of the model, then all the coefficients belonging to this group will be zero. Since group 4 corresponds to the fourth variable (x_4), and contains coefficients of all rules containing x_4, all such rules are left out of the model. Thus, we are left with a complement of the intersection of the groups with group 4, and obtain a subset of rules that does not contain variable 4. The ℓ_1 regularization term here produces general within-group sparsity among all rules to select the most significant rules. Therefore a larger value of λ_2 will induce greater group-level sparsity in the resulting coefficients, $\hat{\boldsymbol{\beta}}$, which, for the selection of groups $\mathcal{G}$ proposed by us will mean greater degree of variable sparsity in the extracted rule set.

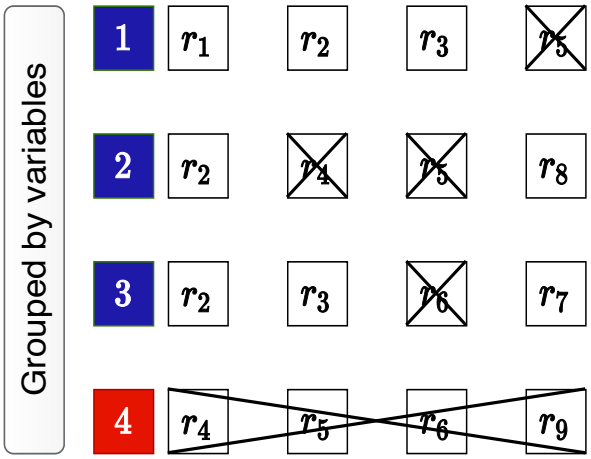


Figure 2.3: Example of variable-sparse structure induced by overlapping group lasso regularization in SURVFIT. When the group corresponding to variable 4 (red) is left out, it zeroes out all the the coefficients of all rules containing variable 4 (crossed out).

2.3.5 Optimization Strategy

The challenge to solve the formulation (2.7) is that the overlapping group structure introduces interdependency among the decision variables of the optimization problem. For non-overlapping groups [59], efficient algorithms that depend on the separability of variables, such as block coordinate descent can be applied [62]. To overcome this challenge, we reformulate this problem as a second-order cone program (SOCP) in Section 2.3.5, and use the interior point method to solve it optimally. However, this strategy will increase the problem size and therefore the computational cost. Alternative methods that may be more efficient include proximal methods, where a key challenge is to develop efficient solutions of the proximal operator. For instance, [65] proposed a smoothing proximal gradient method where a smooth approximation of the overlapping group lasso norm (2.6) and the gradient of this approximation are derived. This smoothing strategy enables a fast iterative shrinkage-thresholding algorithm (FISTA) [66] to solve the problem. [67] proposed the FOGLASSO algorithm using an approximate dual of the proximal operator of the overlapping group lasso norm and its

solution. We discuss their algorithm briefly in the following section. [65] showed that the objective function of their semi-smooth approximation converges to the optimal solution, and [67] observed that though FOGLASSO lacks a convergence guarantee, their algorithm almost always converges to the optimal solution in practice. However, neither study the sparsity structure of the solution they obtained in comparison to the sparsity structure of a solution that does not employ approximations. An understanding of the solution structure obtained is critical in high dimensional problems with multiple optimal solutions where the goal is not just to obtain a solution with an optimal value but also to minimize the number of nonzero coefficients in the solution. [65] showed that the objective function of their semi-smooth approximation converges to the optimal solution, and [67] observed that though FOGLASSO lacks a convergence guarantee, their algorithm almost always converges to the optimal solution in practice. However, neither study the sparsity structure of the solution they obtained in comparison to the sparsity structure of a solution that does not employ approximations. An understanding of the solution structure obtained is critical in high dimensional problems with multiple optimal solutions where the goal is not just to obtain a solution with an optimal value but also to minimize the number of nonzero coefficients in the solution.

SOCP optimization.

In what follows, we cast the formulation (2.7) as a second-order cone program (SOCP). First, we introduce a variable, $z_i = \min(0, t(\mathbf{x}_i) - t_i), \forall i \in \{i \mid \delta_i = 0\}$. Then, we linearize the first regularization term in (2.7) by introducing two new variables, $\boldsymbol{\beta}^+, \boldsymbol{\beta}^-$ such that $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$, $|\boldsymbol{\beta}| = \boldsymbol{\beta}^+ + \boldsymbol{\beta}^-$, and $\boldsymbol{\beta}^+, \boldsymbol{\beta}^- \geq 0$. Third, to deal with the ℓ_2 overlapping group norm, we introduce new variables $\mathbf{y} \in \mathbb{R}^P$ such that

$$\mathbf{y}_p \geq \|\boldsymbol{\beta}_{G_p}\|_2 \quad \forall p \in \{1, \dots P\}. \tag{2.8}$$

The reformulated problem can be written as

$$\min_{\boldsymbol{\beta}^+,\boldsymbol{\beta}^-,\mathbf{z},\mathbf{y}} \|\mathbf{A}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) - \mathbf{t}\|_2^2 + \gamma\|\mathbf{z}\|_2^2 + \lambda_1 \cdot (\boldsymbol{\beta}^+ + \boldsymbol{\beta}^-) + \lambda_2 \sum_{p=1}^{P} q_p \mathbf{y}_p, \tag{2.9}$$

with the constraints,

$$\boldsymbol{\beta}^+, \boldsymbol{\beta}^- \geq 0, \tag{2.10}$$

$$\mathbf{z} \leq 0, \tag{2.11}$$

$$\mathbf{z} \leq \mathbf{B}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) - \mathbf{c}, \tag{2.12}$$

$$\mathbf{y}_i \geq \|(\boldsymbol{\beta}^+_{G_i} - \boldsymbol{\beta}^-_{G_i})\|_2 \quad \forall i \in \{1, \dots P\}. \tag{2.13}$$

Here $\mathbf{A}$ is the rule data matrix corresponding to event observations, $\mathbf{B}$ is the rule data matrix corresponding to censored observations and, $\mathbf{t}$ and $\mathbf{c}$ are the event times and censor times respectively, and (2.13) is a second-order cone constraint. This problem can now be solved by standard SOCP solvers. We use CPLEX in our experiments.

FOGLASSO.

Another algorithm that can solve the formulation approximately is the FOGLASSO algorithm [67] implemented in the **SLEP** [68] MATLAB package. Here we briefly introduce some details of the FOGLASS algorithm for the sake of completeness. FOGLASSO uses an accelerated proximal gradient method where the coefficient estimates at each iteration are, $\boldsymbol{\beta}_{i+1} = \pi^{\lambda_1/L}_{\lambda_2/L}(s_i - \frac{1}{L}l'(s_i))$, where s_i is the affine combination of the current and previous estimates $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_{i-1}$ as used in FISTA [66], L_i is an appropriate constant determined via backtracking line search by the Armijo-Goldstein condition [69], and $\pi(\cdot)$ is the proximal operator of the non-smooth regularization term, $\phi^{\lambda_2}_{\lambda_1}(\boldsymbol{\beta}) = \lambda_1 \cdot \|\boldsymbol{\beta}\|_1 + \lambda_2 \cdot \sum_{p=1}^{P} q_p \|\boldsymbol{\beta}_{G_p}\|_2$. Then, the main optimization problem of the proximal operator is derived to be

$$\pi^{\lambda_2}_{\lambda_1}(\mathbf{v}) = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^P} \Big\{ g^{\lambda_2}_{\lambda_1}(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta} - v\|^2 + \phi^{\lambda_2}_{\lambda_1}(\boldsymbol{\beta}) \Big\}. \tag{2.14}$$

FOGLASSO uses an efficient computational algorithm to solve this proximal operator by reformulating it as a smooth and convex dual problem. A pre-processing step is developed

to identify many zero groups, which reduces the complexity of the optimization problem. However, a trade-off is that their proximal operator solution is inexact, and it is stated that the optimal convergence rate is not guaranteed, though the algorithm works well in practice.

Scalability and Computational Analysis

Here we provide an in-depth analysis of the computational complexity of SURVFIT as well as the existing baseline methods in the literature.

The package **pre** uses the Cox-lasso formulation to extract rules. The time complexity for Cox-regression is $\mathcal{O}(NK^2)$ while that of the Cox-lasso is $\mathcal{O}(NK)$ [70]. However, there are significant computational challenges to optimizing the Cox partial likelihood loss function. The partial likelihood does not naturally decouple over individuals or subsets of individuals, therefore when regularized with a non-smooth term such as the overlapping group lasso, first order proximal gradient descent methods like FOGLASSO, prox-Grad and alternating direction method of multipliers [71] cannot be used. This also means that the stochastic gradient-based optimization methods are not suitable for the task [72]. We have not come across any works that have addressed these computational challenges in our review. One approach could be to use the standard Newton-Ralphson second order scheme [71], however, this is not practical for even medium-sized problems because it involves inverting large matrices at each iteration, which itself has complexity $\mathcal{O}(K^3)$, infeasible to solve our problem of rule selection since we start with a large number of rules. Our approach of using a quadratic loss function instead of Cox-partial log-likelihood loss skirts this issue, allowing us to use efficient first-order and second-order optimization schemes (2.7).

In this paper, we have discussed two different optimization schemes, SOCP and FOGLASSO to solve Problem (2.7). Using the quadratic loss function instead of the Cox partial likelihood loss allows us to formulate the problem as a second-order cone program and solve it using CPLEX solver. The solver uses the barrier interior point method (IPM), a second-order method known to converge in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ iterations, where ϵ is the accuracy at convergence. The per-iteration time complexity of IPM [71] in our case is $\mathcal{O}\big(K^2(N + \sum_{p\in 1\cdots P} |G_p|)\big)$.

First-order proximal gradient methods such as FOGLASSO [68] and Prox-Grad [65] have been proposed to efficiently solve the problem of convex loss functions regularized with an overlapping group lasso loss. The algorithms take more iterations to converge than the second-order IPM algorithm, i.e $\mathcal{O}(\frac{1}{\epsilon})$ iterations. However, the per iteration complexity is lower by orders of magnitude thereby reaching convergence faster. For Prox-Grad, the per-iteration complexity is $\mathcal{O}(NK + \sum_{p\in 1\cdots P} |G_p|)$ [65]. In practice, IPM is found to be more accurate than proximal methods though proximal methods are more efficient and scalable for large-scale problems. Thereby, we see that applying an overlapping group regularization term to induce variable sparsity for the standard Cox partial likelihood loss has a time complexity of at least $\mathcal{O}(NK^3)$ while our formulation can be solved by first-order methods in $\mathcal{O}(NK+\sum_{p\in 1\cdots P} |G_p|)$ and second order methods in $\mathcal{O}\big(K^2(N+\sum_{p\in 1\cdots P} |G_p|)\big)$. This is much more efficient in high-dimensional problems such as rule extraction where often $K >> N$, i.e., the number of rules is much greater than the number of observations.

2.3.6 Interpretability of the Rules

Most rule learning methods in the literature of machine learning only concern rule discovery, namely the identification of important rules from data such as Rulefit [20] and **pre** [25]. For instance, a rule suggests an interaction among a set of variables, and the essence of an interaction is that the variables give a greater effect when combined than taken individually. A rule-learning algorithm may generate a set of rules, but it cannot prove that the interactions are genuine. Here, we further develop a rule analysis framework that employs a combination of statistical methods such as survival data analysis, hypothesis testing, and regression analysis to evaluate the significance and to better understand the implications of the discovered rules in various contexts.

Statistical Testing.

We use statistical testing to evaluate whether the extracted rules are significantly associated with the outcome. Specifically, we analyze whether the subjects endorsing each rule have a

significantly higher or lower risk of onset of the event as compared to subjects not endorsing the rule. For this goal, the Kaplan-Meier curve is used to study the separation of the survival functions of the groups of observations defined by each rule. We also employ the log-rank test [73], which is a hypothesis testing method used to examine differences in risk of event occurrence between the two groups.

Decomposition Analysis.

Decomposition analysis is used to examine the rules to see if the interactions among the variables are genuine. Basically, we decompose the rule by removing one variable at a time and evaluating the impact of this removal, i.e., by statistically evaluating the difference using the Kaplan-Meier curve and log-rank tests. If the removal of a variable is found to have little impact on the overall significance of the rule, which is possible since the rule learning algorithm uses a greedy predictive metric to guide the rule selection process, then we should trim the rule by removing this variable. On the other hand, the variable which has the greatest impact on the significance of the rule is said to be the dominant or critical variable of that rule. And if the combination of two or more variables has a much higher association with the risk than either of the variables taken individually, then the interaction of those two variables is highly significant in predicting the risk. Those are possible scenarios the decomposition analysis could shed light on and reveal more understanding of the rules and their constituent variables.

Sensitivity Analysis.

Besides the combinatorial characteristic of the rules, cutoff values of the constituent variables are also essential information in defining the interactions among the variables. Sensitivity analysis is conducted to evaluate how sensitively the statistical significance of the rule depends on the cutoff values of the variables used in the rule. We study the change in the odds ratio of the two groups defined by the rule as the cutoff value of a variable in a rule changes. The odds ratio (OR) [74] is the ratio of the probability of event occurring in the

subgroup endorsing the rule to the ratio of the event occurring in the other subgroup. This sensitivity analysis would reveal different scenarios for the cutoff values as well, e.g., there is sometimes indeed the best cutoff value for a factor, with a cutoff value that maximizes the statistical significance of the rule, and around the best value there is either a sharp or smooth descending slope. For some other variables, there seems to be a range of cutoff values that are equally good. Thus, sensitivity analysis could reveal unique insights regarding the variables and the rules that engage them.

2.4 Numerical Experiments

In this section we conduct numerical experiments to evaluate the proposed SURVFIT method, compare the solutions produced by the two optimization strategies described in Section 2.3.5, and compare SURVFIT with the baseline approach that uses a regularized cox regression model and survival random forest. One of our goals is to demonstrate the variable sparsity property of SURVFIT. Therefore, the rules extracted with (2.7) and without (4.2) doubly sparse regularization, and their decomposition, and sensitivity analysis, are also presented. Predictive and variable selection performance for each of the models are evaluated over 100 repetitions of the same experimental setup, with 80/20 partitions of the data for training/testing.

2.4.1 Evaluation Criteria

The following measures are used to rank, and evaluate the significance of the rules extracted by SURVFIT.

1. Importance. We rank the rules obtained by our model by the importance measure, $\mathrm{I}(r)$ which is defined as

$$\mathrm{I}(r) = \hat{\beta}_r \sqrt{s(r)(1 - s(r))}, \tag{2.15}$$

where $s(r)$ is the support and $\hat{\beta}_r$ is the coefficient estimate of rule r.

2. Support. The support of a rule, $s(r)$ is defined as the fraction of total observations that endorse the rule,
$$s(r) = \frac{\sum_{i=1}^{N} r(\mathbf{x}_i)}{N}. \tag{2.16}$$

3. Odds Ratio. The odds ratio is the ratio of the odds of event occurrence in the data endorsing the rule to the odds of event occurrence in data not endorsing the rule.

Furthermore, the following measures are used to evaluate and compare the performance of our method with baseline models.

1. Concordance index (C-Index): Harell's concordance index [75] is used to estimate and compare the predictive performance of survival models. The c-index estimates the probability that in randomly selected pair of test subjects, the subject with the earlier event occurrence has an earlier model prediction of event time. Therefore a completely random prediction will achieve a c-index of 0.5.

2. False positive rate (FPR): The FPR is a measure of the models capability to select a sparse set of significant variables. It is defined as the ratio of the number of incorrect (or noisy) variables selected by the model to the total number of variables.

2.4.2 Simulation Study

We simulate a dataset consisting of $N = 2000$ observations with $P = 60$ variables using an approach similar to the one adopted in [20]. The event times (2.17) are simulated such that only 7 variables ($x_1 - x_7$) affect the target, another 7 variables ($x_{18} - x_{14}$) are correlated in varying degrees to the first 7 variables, and the remaining 46 variables are purely noise. The response variable t_i for each input $\mathbf{x}_i$ is taken to be

$$t_i = F(\mathbf{x}_i) + \epsilon_i, \tag{2.17}$$

where the function $F(x)$ is defined as

$$F(x) = 4\prod_{j=1}^{3}(-(1-x_j)^2) - 0.55 * \exp\{-2(x_4 - x_5)\} + 1.75 * \sin(x_6 - x_7) \tag{2.18}$$

and $\epsilon \sim N(0, \sigma^2)$, where σ^2 is chosen to keep a signal to noise ratio of 3. The response $\mathbf{t}$ is then scaled to make sure it is positive. The parameters of this function are chosen in a way to obtain an approximately equal representation of each predictive variable in the exhaustive rule list. We assume that 35% of the simulated data is censored. To account for this, the event times of a random subset consisting of 35% of the data are multiplied with a uniform random variable to simulate the censored times. The remaining 65% of the data is assumed to be complete. To simulate each of the correlated variables in $\{x_8 \dots x_{14}\}$, we first sample a correlation $\rho \in (0, 1)$ from a uniform distribution by which it is correlated to the corresponding original variable in $\{x_1 \dots x_7\}$, then we choose $x = \rho \cdot \sigma_{x^*} x + \sqrt{1-\rho^2} \cdot \sigma_x x^*$, where x is the correlated variable in $\{x_8 \dots x_{14}\}$ and x^* is the residual of a least-squared regression between x and its corresponding original variable in $\{x_1 \dots x_7\}$. The response simulation model (2.18) involves explicit interactions between (x_1, x_2, x_3) in the first term, (x_4, x_5) in the second term, and between (x_6, x_7) in the third term.

2.4.3 Synthetic Data Results

Comparison of the SOCP algorithm with FOGLASSO.

Fig. 2.4 compares the number of variables included in the top ranked rules extracted at various values of λ_2 for the SOCP and FOGLASSO methods. Firstly, both algorithms show the effectiveness of the SURVFIT formulation to produce a doubly-sparse set of rules which are also sparse in the number of variables involved in the rules, i.e., it is observed that the extracted rules contain fewer variables when λ_2 is increased. Secondly, we observe that compared with our SOCP algorithm, FOGLASSO leads to rules with more variables. It is worth mentioning that in our experiments, solving the problem using proximal smoothing algorithm introduced by [65] leads to a selection of rules which had even more variables than FOGLASSO. This difference in the structure of the solutions under these algorithms may be attributed to the approximate nature of the algorithm used to solve the proximal operator in FOGLASSO, and the smoothing approximation of the overlapping group norm (2.6) used by

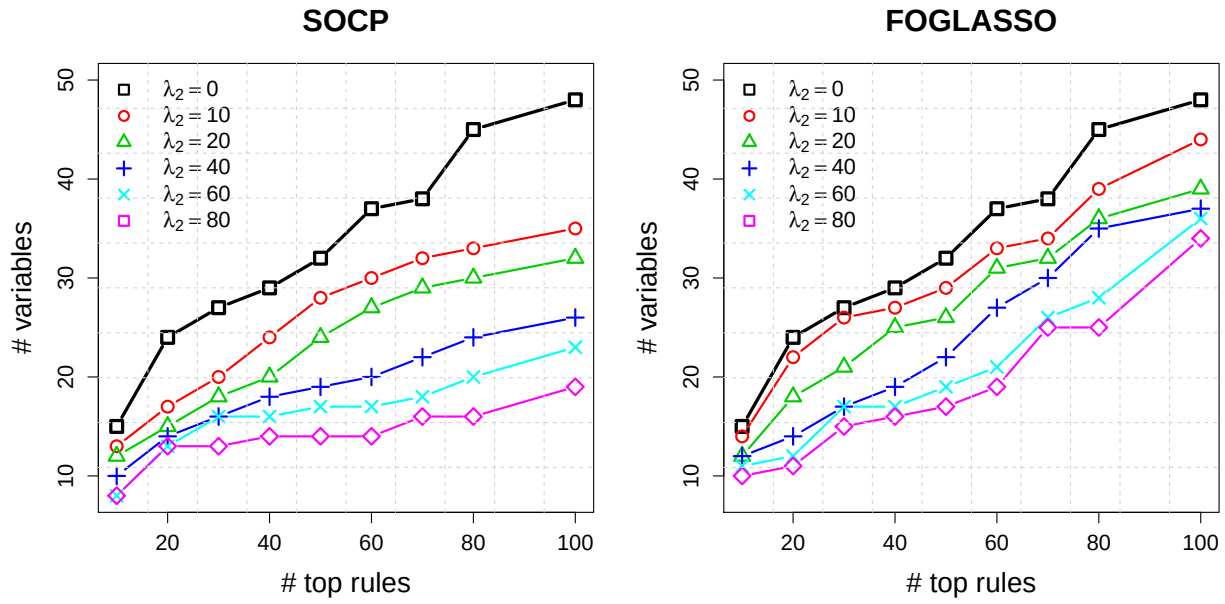


Figure 2.4: Number of variables included in the top rules extracted at various values of λ_2 for the SOCP (left) and FOGLASSO (right) optimization methods.

[65] which lead to a less sparse solution in terms of groups i.e variable selection, a trade-off for their computational advantage.

Quality and efficiency of rule extraction.

The hyperparameters $\lambda_1 = 5$ and $\lambda_2 = 20$ are chosen after cross-validation for the following analysis. The SURVFIT algorithm then extracts the highly predictive rules which are also sparse in the number of variables among them. SURVFIT, like other rule models based on sparse regularization, produces a rank for each rule. One way to compare different methods is to see the quality of highly ranked rules, i.e if noisy or low quality rules have high rank, the algorithm is less useful. The rules are ranked based on the *importance* measure introduced in Section 2.4.1. In our experiment, we compare the top 8 rules which are extracted by SURVFIT (in Table 2.2) and the rules extracted by ℓ_1 penalized optimization without the variable sparsity penalty (in Table 2.1). One main interest of this numerical study is to see

if SURVFIT could detect significant rules without falsely introducing noisy variables. Table 2.2 presents the top 8 rules extracted via SURVFIT which involve 7 variables, x_1, x_2, $x_4 - x_7$ and a singular false positive, x_{11}. (The significant variable x_3 is not picked in the top 8 rules though it does show up in the total list of significant rules with non-zero coefficients). This is consistent with the ground truth that has been used in the synthetic data generation as shown in (2.17). A remarkable observation is that SURVFIT is resilient to the noise in the data which was designed for variables $x_8 - x_{14}$ to be statistically correlated with $x_1 - x_7$. To further demonstrate this point, Table 2.1 lists the top 8 rules extracted without doubly-sparse regularization, i.e. only employing ℓ_1 norm penalty to obtain sparsity in the cardinality of rules. In contrast to SURVFIT we can observe that the rules in Table 2.1 contain a total of 15 variables, of which many are false positives. Thus, doubly-sparse regularization used in SURVFIT is effective to recover the true variables and genuine interactions in the data-generating process. Fig. 2.4 shows the number of variables involved in the top rules returned by SURVFIT for different values of λ_2. Increasing the value of the variable-sparsity parameter, λ_2 leads to rules which are more sparse in the number of variables. For instance, at $\lambda_2 = 0$, the top 10 rules involve 16 variables, while for $\lambda_2 = 80$, the top 10 rules involve only 7 variables. The difference is even starker when comparing the top 100 rules: there are close to 50 variables at $\lambda_2 = 0$, while at $\lambda_2 = 80$, the top 100 rules only contain 16 variables. We conclude that our proposed SURVFIT approach is able to extract a doubly sparse set of rules involving only the most significant variables in the data.

Rule analysis.

Table 2.2 not only presents the top 8 rules extracted via SURVFIT, but also their decomposition, p values of the log-rank test, and support (2.4.1) of each of these rules. Also, Fig. 2.5 shows the Kaplan-Meier survival curves for these rules. We observe from the Kaplan-Meier curves that, endorsement of rules 1-6 and rule 8 is associated with higher survival rates, while endorsement of rule 7 is associated with lower survival rate. As we show in the following discussion, these rules recover the variable effects and interactions on mortality encoded

in the data generation process. The decomposition analysis (besides the p-values shown in Table 2.1, the decomposition curves for the rules are also plotted in Fig. 2.6) also reveals interesting insights into the data. For instance, on rule 2 we can see that there is a real interaction between x_4 and x_5, as removal of either hugely impacts the significance of the rule. This is consistent with the ground truth model as the way the two variables are incorporated into the data-generating mechanism is through the functional $\exp\{-2(x_4 - x_5)\}$. A similar observation holds true for rule 3 and its decomposition analysis reveals the interaction between x_6 and x_7. The decomposition analysis on rule 4 is also insightful, as it actually shows that with x_5 alone the significance of the rule is stronger, indicating that there is no synergistic interaction between x_1 and x_5 so there is no need to keep variable x_1 in the rule. This shows that rule 4 contains noise resulting from the greedy nature of the tree-growing algorithm used. While the double-sparsity enforced by SURVFIT aims to reduce this noise, we also need decomposition analysis to further filter out any residual noise. A similar conclusion holds true for the decomposition analysis on rule 5, rule 6, and rule 7. We could still keep some significant, but not explicitly interacting variables in the rules, e.g., x_2 in rule 5 and x_1 in rule 6, since it is hard to call these two rules interrupted by noise, given that x_2 and x_1 contribute to the overall significance of the rules though their contributions are quite marginal. Here there is no obvious contradiction with the ground truth model since all the variables are truly involved in the ground truth, although not in explicit interaction terms, an implicit interaction exists.

Next we analyze the sensitivity of the cutoff values of the variables. Fig. 2.7 shows how the overall significance of the rules, as measured by the odds ratio and its 95% confidence interval changes with change in the cutoff value of the variables in the rules. For instance, in the sensitivity analysis of Rule 1, we see that increasing the cutoff value of x_5 reduces the odds ratio, meaning that at higher values of x_5 the probability of event occurrence falls off. This is consistent with both the Kaplan Meier curve associated with rule 1 as well as the data-generating mechanism. The sensitivity analysis of rule 2 shows how the odds ratio would change with change in cutoff of each of x_4 and x_5, while keeping the other cutoff constant. At

Table 2.1: Top 8 Rules Identified without double sparsity penalty (4.2) and Corresponding Log-Rank p-Values

Rule ID	Rule	p Value	Support (%)
1	$x_4 > 0.532$ AND $x_{49} > 0.033$	$8e-13$	37.05
1a	$x_4 > 0.532$	$4e-15$	38.3
1b	$x_{49} > 0.033$	0.01	96.6
2	$x_5 > 0.47$ AND $x_{10} <= 0.705$	$3e-19$	42.25
2a	$x_5 > 0.47$	$3e-16$	46
2b	$x_{10} <= 0.705$	$3e-04$	92
3	$x_2 <= 0.309$ AND $x_5 > 0.5$ AND $x_{11} <= 0.289$	$6e-12$	6.40
3a	$x_5 > 0.5$ AND $x_{11} <= 0.289$	$2e-14$	18.40
3b	$x_2 <= 0.309$ AND $x_{11} <= 0.289$	$2e-12$	15.90
3c	$x_2 <= 0.309$ AND $x_5 > 0.5$	$1e-13$	14.90
4	$x_2 <= 0.45$ AND $x_6 > 0.486$ AND $x_{57} <= 0.892$	$3.5e-18$	23.25
4a	$x_6 > 0.486$ AND $x_{57} <= 0.892$	$2.5e-17$	44.05
4b	$x_2 <= 0.45$ AND $x_{57} <= 0.892$	$5e-12$	50.60
4c	$x_2 <= 0.45$ AND $x_6 > 0.486$	$2e-18$	23.55
5	$x_1 <= 0.25$ AND $x_6 > 0.442$ AND $x_{11} <= 0.346$	$6e-13$	6.90
5a	$x_6 > 0.442$ AND $x_{11} <= 0.346$	$6e-16$	25.80
5b	$x_1 <= 0.25$ AND $x_{11} <= 0.346$	$1e-09$	15.25
5c	$x_1 <= 0.25$ AND $x_6 > 0.442$	$1e-13$	13.45
6	$x_6 > 0.48$ AND $x_7 > 0.48$ AND $x_{30} <= 0.893$	$2e-22$	20.85
6a	$x_7 > 0.48$ AND $x_{30} <= 0.893$	$3e-10$	46.20
6b	$x_6 > 0.48$ AND $x_{30} <= 0.893$	$2e-17$	44.90
6c	$x_6 > 0.48$ AND $x_7 > 0.48$	$3e-22$	21.00
7	$x_4 <= 0.494$ AND $x_9 <= 0.553$ AND $x_{13} > 0.549$	$3e-12$	10.35
7a	$x_9 <= 0.553$ AND $x_{13} > 0.549$	$2e-11$	18.60
7b	$x_4 <= 0.494$ AND $x_{13} > 0.549$	$9e-09$	12.60
7c	$x_4 <= 0.494$ AND $x_9 <= 0.553$	$4e-15$	44.30
8	$x_4 <= 0.323$ AND $x_{24} > 0.886$ AND $x_{28} > 0.0043$	$3e-10$	61.45
8a	$x_{24} > 0.886$ AND $x_{28} > 0.0043$	0.2	98.00
8b	$x_4 <= 0.323$ AND $x_{28} > 0.0043$	$2e-11$	62.65
8c	$x_4 <= 0.323$ AND $x_{24} > 0.886$	$4e-15$	61.60

Table 2.2: Top 8 Rules Identified with double sparsity penalty (2.7) and Corresponding Log-Rank p-Values. The final rules selected after decomposition analysis are highlighted in gray.

Rule ID	Rule	p Value	Support (%)
1	$x_5 > 0.687$	$7.5e-15$	21.5
2	$x_4 <= 0.52$ AND $x_5 > 0.395$	$7e-25$	32.80
2a	$x_4 <= 0.52$	$8e-15$	60.20
2b	$x_5 > 0.395$	$3e-14$	54.40
3	$x_6 > 0.479$ AND $x_7 > 0.48$	$3e-23$	21.0
3a	$x_6 > 0.479$	$2e-17$	45.25
3b	$x_7 > 0.48$	$3e-10$	46.55
4	$x_1 < 0.4$ AND $x_5 > 0.71$	$2e-9$	8.05
4a	$x_1 < 0.4$	$2e-6$	44.80
4b	$x_5 > 0.71$	$1e-10$	18.25
5	$x_2 <= 0.45$ AND $x_6 > 0.485$	$6e-18$	23.55
5a	$x_2 <= 0.45$	$4e-10$	51.00
5b	$x_6 > 0.485$	$2e-17$	44.55
6	$x_1 <= 0.52$ AND $x_6 > 0.625$	$2e-18$	18.20
6a	$x_1 <= 0.52$	$4e-4$	58.10
6b	$x_6 > 0.625$	$5e-17$	30.75
7	$x_1 >= 0.33$ AND $x_2 > 0.65$ AND $x_7 > 0.428$	$4e-11$	9.30
7a	$x_2 > 0.65$ AND $x_7 > 0.428$	$2e-04$	14.95
7b	$x_1 >= 0.33$ AND $x_7 > 0.428$	0.9	33.90
7c	$x_1 >= 0.33$ AND $x_2 > 0.65$	$2e-12$	17.35
8	$x_2 < 0.31$ AND $x_5 > 0.5$ AND $x_{11} <= 0.288$	$6e-12$	6.4
8a	$x_5 > 0.5$ AND $x_{11} <= 0.288$	$2e-14$	18.40
8b	$x_2 < 0.31$ AND $x_{11} <= 0.288$	$2e-12$	15.90
8c	$x_2 < 0.31$ AND $x_5 > 0.5$	$1e-13$	14.90

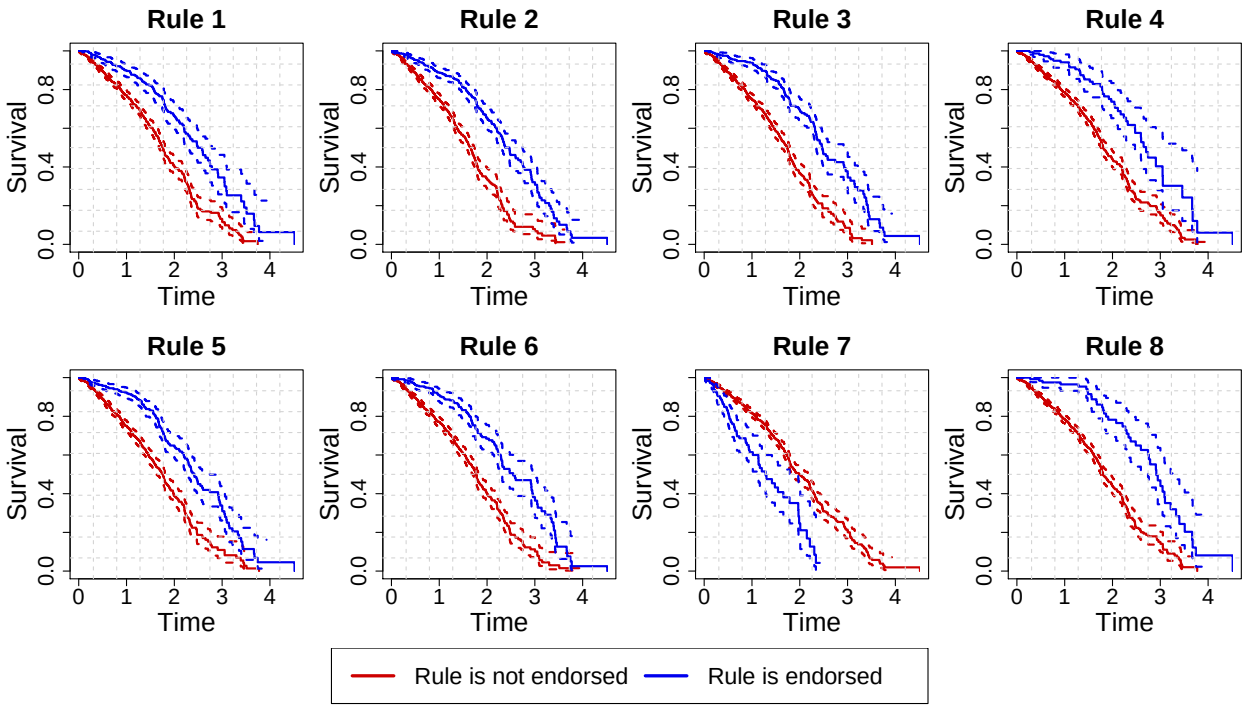


Figure 2.5: Kaplan–Meier survival curves with 95% confidence intervals for each rule in Table 2.2

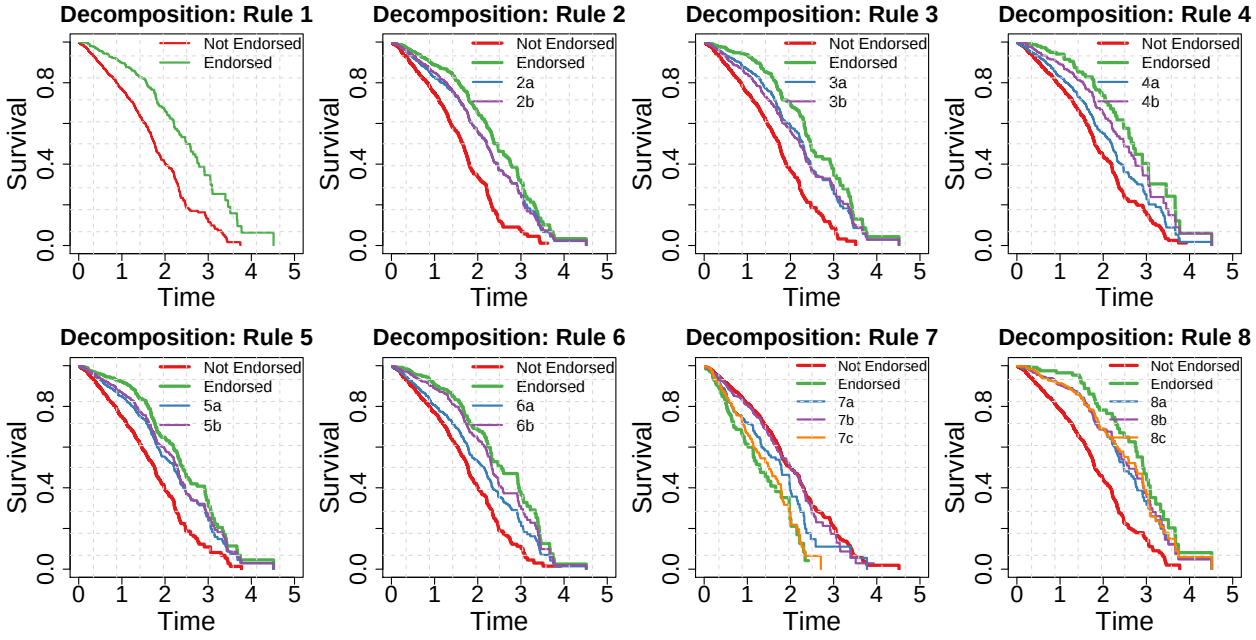


Figure 2.6: Decomposition analysis for each rule in Table 2.2

smaller cutoff values of x_4, the odds ratio is greater than 1 signifying higher risk levels while at higher cutoffs of x_5, the cutoff is lower than 1 signifying lower risk. This is again consistent with the ground truth since the time of event occurrence depends on $-\exp(-2(x_4 - x_5))$, and a larger value of x_4 would decrease and a greater value of x_5 would increase the value of this term and hence decrease and increase the time of event occurrence, respectively. The sensitivity analysis of the other rules can also be interpreted in this context and seen to be consistent with our knowledge of the ground truth. Thus, the sensitivity analysis helps us understand how each variable affects the event risk under different conditions.

Comparison of predictive and sparsity performance of SURVFIT with standard survival models.

We compare the predictive performance of the SURVFIT algorithm to standard survival analysis methods such as random survival forests implemented in **randomForestSRC** [56],

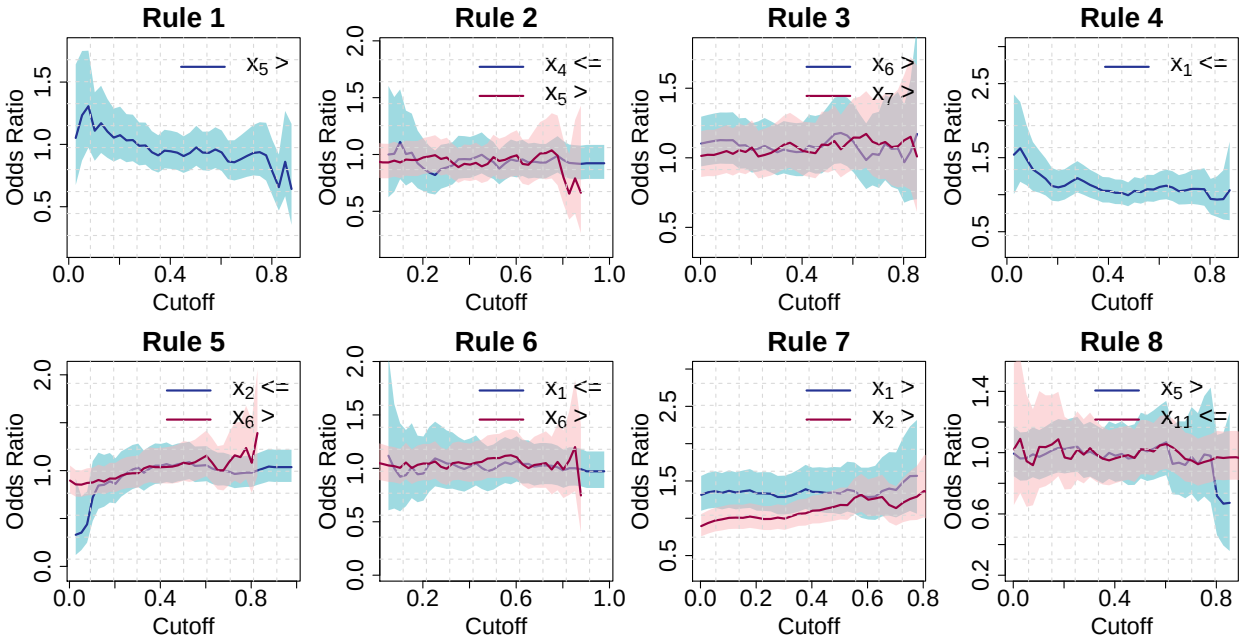


Figure 2.7: Sensitivity analysis of rules from Table 2.2

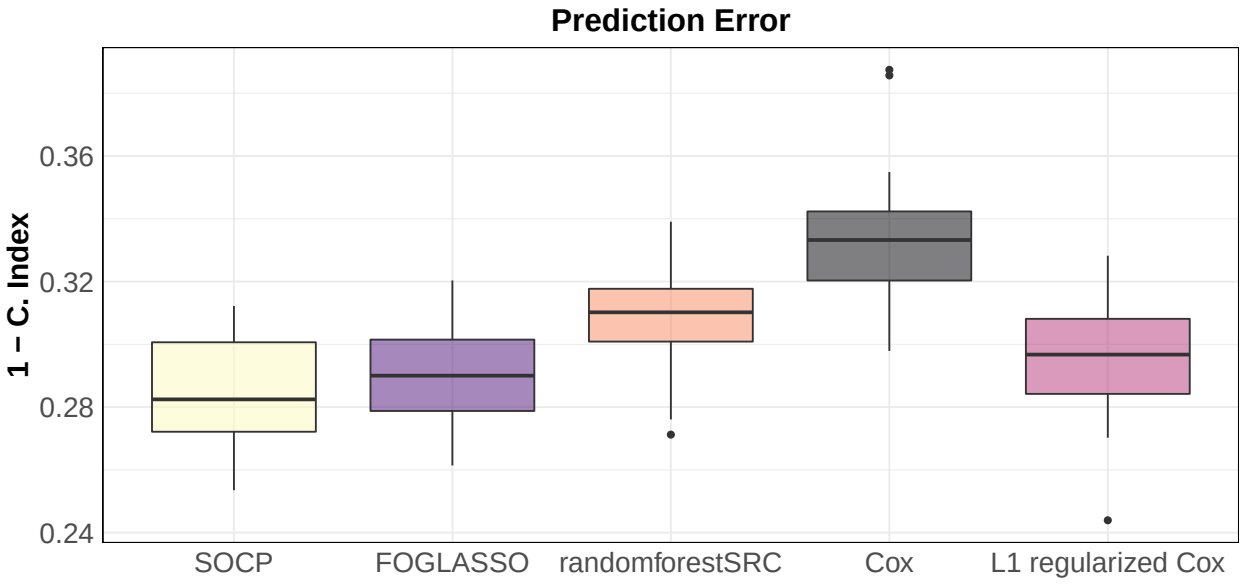


Figure 2.8: Comparison of prediction error of SURVFIT with standard survival analysis methods on synthetic data

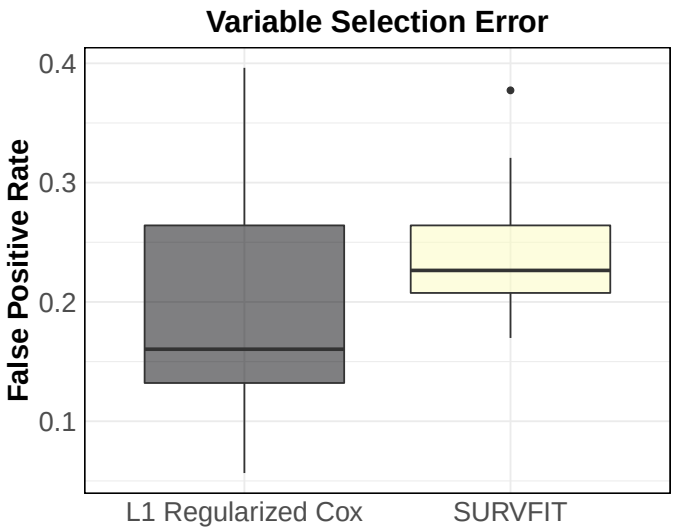


Figure 2.9: Comparison of sparsity performance of Regularized Cox and SURVFIT model on synthetic data

Cox regression [61], and regularized Cox regression [58] using the concordance index (C Index) metric (2.4.1). The C Index for SURVFIT is calculated through the times-of-event occurrence estimated by SURVFIT as shown in Section 2.4.1. Fig. 2.8 shows box plots of the estimates of prediction error (1 - C Index) obtained over 100 independently sampled replicates. The following performance evaluation procedure has been adopted: first we sample a training set of 1600 observations, and then an independent test set of 400 observations. The different models are then trained on the training dataset and the reported performance evaluation is based on the test set. This procedure is repeated 100 times to obtain the C-Index estimates for each of the different methods. It can be observed that the prediction errors of SURVFIT, when considering rules with all non-zero coefficients is lower than the prediction errors of other methods. To compare the error rates of the different methods, we use the paired Wilcoxon rank sum test on our C-Index estimates. For each pair of methods, we perform the following test:

Null hypothesis: $\text{C-Index}_1 = \text{C-Index}_2$

Alternative hypothesis: C-Index$_1$ $\neq$ C-Index$_2$

The p-Values from each of these pairwise tests are provided in Table 2.3. It can be seen that the performance of the methods are significantly different from each other.

Table 2.3: p-value of pairwise Wilcoxon rank sum test on C-index obtained by each of these methods on synthetic data

	FOGLASSO	randomForestSRC	Cox	L1 regularized Cox
SOCP	0.0002563	9.3e-09	1.86e-09	0.00113
FOGLASSO		1.86e-08	1.86e-09	0.04592
randomForestSRC			1.82e-06	1.3e-07
Cox				1.86e-09

Our goal is not only to get a model that is accurate in terms of prediction but also exhibits sparsity in the number of variables. To do this we compare the false positives in the variables involved in the SURVFIT model with variables involved in the regularized Cox regression model. Fig. 2.9 compares a box plot of the variable selection error (false positive rate, (2)) of the regularized Cox model and SURVFIT obtained over 100 independently sampled replicates. As the figure shows, regularized Cox-regression does slightly better on average than SURVFIT on our synthetic data in terms of variable selection, although the spread of error is higher. The other models like randomForestSRC and Cox regression use all variables in the data, therefore a comparison of variable selection with these models is not meaningful. The true positive rate, i.e., the proportion of correct variables identified is equal to 1 for both models, i.e., both SURVFIT and regularized Cox regression select all of the significant variables.

2.4.4 MIMIC-III Sepsis Data Study

MIMIC-III (Medical Information Mart for Intensive Care) [76] is a comprehensive database comprising anonymized information relating to patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA between 2001 and 2012. The data consists of over 53,000 adult ICU admissions during this time period. In this paper, we utilize a subset of inpatient admissions that were diagnosed with at least one of sepsis, or severe sepsis, or septic shock, which are increasingly severe sepsis conditions. This subset has 2,840 samples in total. Sepsis is a common ailment caused by infections and characterized by whole body inflammation which accounts for 2% of hospitalizations and 25% of ICU bed utilizations in the United States. It is the second leading cause of death among ICU patients, the third leading cause of death worldwide, and the main cause of hospital mortality [77, 78]. Understanding the mortality risk from sepsis would be beneficial for physicians in selecting a more efficient management approach. Several recent studies have focused on predicting mortality risk based on variables related to predisposition [79], pre-existing and co-morbid conditions [80], cytokines and immune system interleukin's [81], and gene expression analysis [82]. A recent study on early sepsis detection by [83] has used heart rate and blood pressure dynamics data. As the mechanism of how these variables impact the mortality risk is known to be complex, we use SURVFIT to extract rules and study the interactions among the variables. Out of the 2,840 patient observations in our dataset, 1,097 (38.6%) are mortal event instances with a record of time of death and the remaining are censored with time of discharge as the censor time. We investigated 78 variables in our analysis, consisting of patient characteristics such as age, race, gender, weight, and clinical history; physiological measurements such as respiratory rate, blood pressure, heart rate, oxygen saturation etc., and summary statistics of physiological measurements and laboratory test results such as blood urea nitrogen, creatinine, white blood cell count, and hemoglobin etc.

Sepsis Survival Results.

We choose $\gamma = 5 \times 10^{-6}$, $\lambda_1 = 50$ and $\lambda_2 = 10$ for our SURVFIT model through cross-validation.

The top 8 rules extracted by SURVFIT are presented in Table 2.4, along with their p-values of the log-rank test, their support, and the results of the decomposition and sensitivity analysis. We obtain a total of 13 significant variables involved in the top 8 rules affecting survival risk. Aspartate-aminotransferase, oxygen saturation (O_2-sat.), Alanine-aminotransferase, arterial-pH, age, heart-rate, Alanine-aminotransferase(tests), diastolic BP (noninv-dia-BP), length of stay, systolic BP (noninv-sys-BP) are the variables associated with sepsis mortality risk. We quantitatively, and descriptively evaluate their interaction effects on mortality. Each of these rules is significant, i.e., as shown in the p values of the log-rank test. The Kaplan–Meier curves of the rules are shown in Fig. 2.10. The Kaplan–Meier curves reveal that the rules 1, 2, 4, 5 and 7 are risk-reducing rules, i.e., patients who endorse these rules have less risk of mortality, and the rules 3, 6 and 8 are risk-increasing rules. Based on the decomposition analysis of the rules, i.e., p-values shown in Table 2.4 and survival curves of decomposition analysis shown in Fig. 2.11, we are able to gain a greater understanding of the nature of the interactions of the variable in each of the rules. For example, in rule 1, while both Aspartate-aminotransferase(mean) and oxygen saturation, O_2-sat.(mean) are significant in predicting the risk, O_2-sat.(mean) is the critical factor due to its lower p value. In rule 2, the interaction between Alanine-aminotransferase (mean) and arterial-pH (mean) is significant in predicting the mortality. Decomposition analysis of rule 3 demonstrates an interaction of heart-rate(sd) and arterial-pH(mean) to be highly significant while the interaction between age and arterial-pH is not. The removal of heart-rate(sd) from rule 4 reduces the rule discrimination ability the most, making it the critical factor. Likewise, in other rules we observe that diastolic blood pressure (noninv-dia-BP), systolic blood pressure (noninv-sys-BP) and length of stay (total LOS) also influence the mortality rate. In rule 8, decomposition analysis reveals that noninv-sys-BP(mean) is the critical

factor of the rule while heart-rate(tests) has no contribution despite being involved in the rule. The literature studying sepsis mortality supports our results, as the variables covered in these rules as well as their cutoff values have been found to be significant in predicting the mortality associated with sepsis. For example, rule 1 suggests that higher saturated oxygen, (O_2-sat(mean) > 92.5) is associated with lower mortality risk. Our findings are consistent with the results found by [84] who reported a lower level of oxygen saturation in non-survivors as compared to survivors, and a value below 78 is associated with increased risk of mortality among patients of septic shock in their experiments. Alanine aminotransferase (alt), and aspartate aminotransferase (ast) are liver enzymes that are biomarkers of abnormal liver functions which is often found in sepsis patients [85]. In rules 1 and 2 we see that a higher Aspartate-aminotransferase and Alanine aminotransferase signifies increased mortality risk, high levels of both enzymes have been found to significant predictors of sepsis-associated liver injury [86, 87] in literature. In rule 6, standard deviation of oxygen saturation (O_2-sat(sd)) is found to be a significant predictor of mortality, and higher deviations are associated with higher mortality. A similar result was found by [88] who investigated spontaneous changes in oxygen saturated in sepsis patients and reported a significantly higher number of severe changes in O_2-sat in non-surviving patients when compared to surviving patients. However, the literature does not discuss the significance of interactions between the variables found in our model.

Sensitivity analysis of some critical factors of the rules are reported in Fig. 2.12. The sensitivity analysis figures show the odds ratio (and 95% confidence interval) of the rules change when the cutoff values of variables are changed while keeping cutoffs of other variables in the rule at the base level. Analysis of O_2-sat. in rule 1 shows that an O_2-sat.(mean) value greater than 90 leads to an odds ratio much lower than 1, and hence decreased mortality risk. A further increase in the O_2-sat. shows a steady increase in risk showing that very high levels of oxygen saturation (above 95%) will increase mortality risk. This analysis is in line with a study by [89] which reports that both abnormally high and low levels of oxygen saturation are associated with increased mortality in patients with suspected sepsis. Our

Table 2.4: Top 8 rules identified with doubly sparse penalty from Sepsis survival data and their decomposition analysis. The final rules selected after decomposition analysis are highlighted in gray.

ID	Rule	p Value	Support
1	Aspartate-aminotransferase (mean)<= 308 AND O_2-sat. (mean)> 92.5	$2e-70$	89.6
1a	Aspartate-aminotransferase (mean)<= 308	$2e-38$	91.62
1b	O_2-sat. (mean)> 92.5	$2e-83$	97
2	Alanine-aminotransferase (mean)< 2778.3 AND arterial-pH (mean)> 7.2 AND O_2-sat. (sd)<= 3.23	$8e-150$	81.83
2a	arterial-pH (mean)> 7.2 AND O_2-sat. (sd)<= 3.23	$3e-148$	81.93
2b	Alanine-aminotransferase (mean)< 2778.3 AND O_2-sat. (sd)<= 3.23	$8e-59$	86.97
2c	Alanine-aminotransferase (mean)< 2778.3 AND arterial-pH (mean)> 7.2	$8e-240$	92.18
3	age > 73.85 AND heart-rate (sd) <= 38.74 AND arterial-pH (mean) > 7.25125	$3e-05$	37.21
3a	heart-rate (sd) <= 38.74 AND arterial-pH (mean) > 7.25125	$3e-240$	92.14
3b	age > 73.85 AND arterial-pH (mean) > 7.25125	$1e-05$	37.39
3c	age > 73.85 AND heart-rate (sd) <= 38.74 > 7.25125	$4e-19$	40.21
4	has.septicshock = F AND Alanine-aminotransferase (tests) > 3.5	$2e-23$	34.78
4a	has.septicshock = F	$3e-17$	50.38
4b	Alanine-aminotransferase (tests) > 3.5	$1e-23$	69.82
5	noninv-dia-BP (mean)> 34.3 AND O_2-sat. (sd)<= 5.8 AND noninv-sys-BP (mean) > 111.5	$4e-32$	25.03
5a	O_2-sat. (sd)<= 5.8 AND noninv-sys-BP (mean) > 111.5	$3e-32$	25.07
5b	noninv-dia-BP (mean)> 34.3 AND noninv-sys-BP (mean) > 111.5	$4e-25$	26.30
5c	noninv-dia-BP (mean)> 34.3 AND O_2-sat. (sd)<= 5.8	$3e-60$	95.21
6	Aspartate-aminotransferase (mean)<= 2585 AND O_2-sat. (sd)> 3.1	$5e-42$	13.6
6a	Aspartate-aminotransferase (mean)<= 2580	$5e-33$	98.97
6b	O_2-sat. (sd)> 3.1	$2e-51$	14.01
7	total-los> 0.52 AND heart-rate (tests) s> 478 AND arterial-pH (mean)> 7.25	$1e-37$	12.07
7a	heart-rate (tests) s> 478 AND arterial-pH (mean)> 7.25	$1e-37$	12.07
7b	total-los> 0.52 AND arterial-pH (mean)> 7.25	$5e-290$	90.59
7c	total-los> 0.52 AND heart-rate (tests) s> 478	$2e-37$	12.11
8	heart-rate (tests) > 7 AND noninv-sys-BP (mean)<= 99.61	$2e-27$	8.91
8a	heart-rate (tests) > 7	1	99.4
8b	noninv-sys-BP (mean)<= 99.61	$2e-30$	9.08

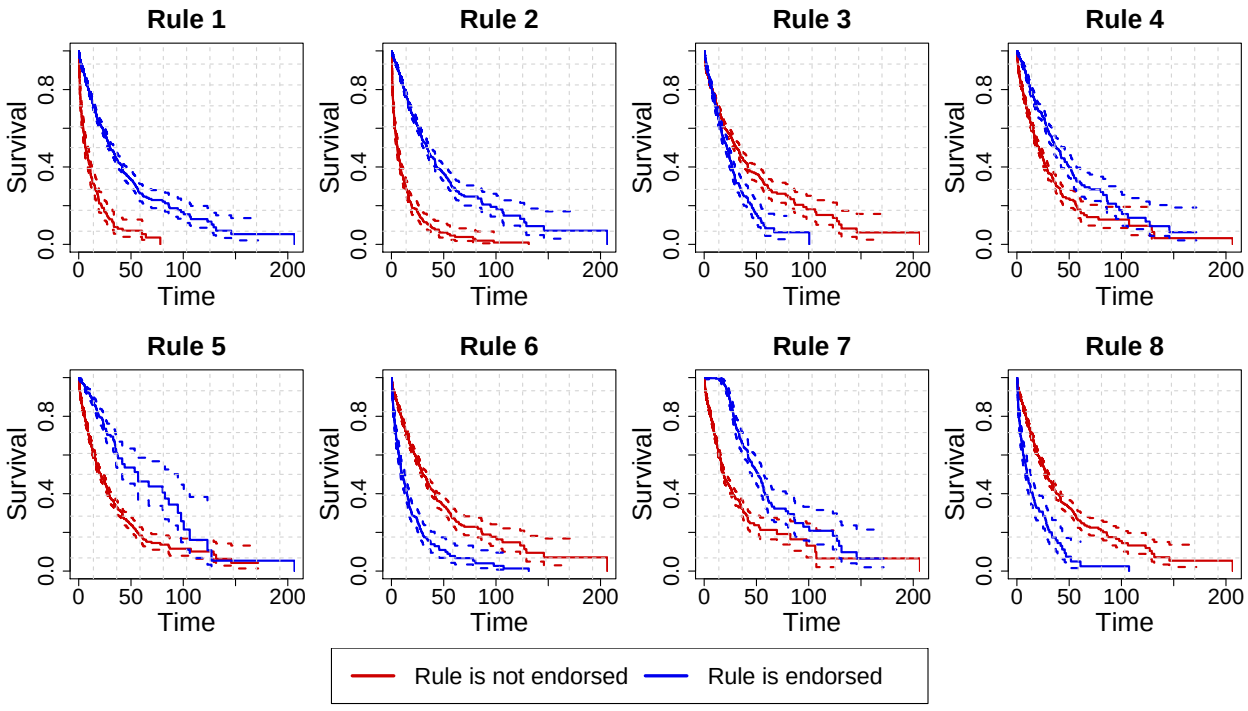


Figure 2.10: Kaplan–Meier survival curves with 95% confidence intervals for rules in Table 2.4

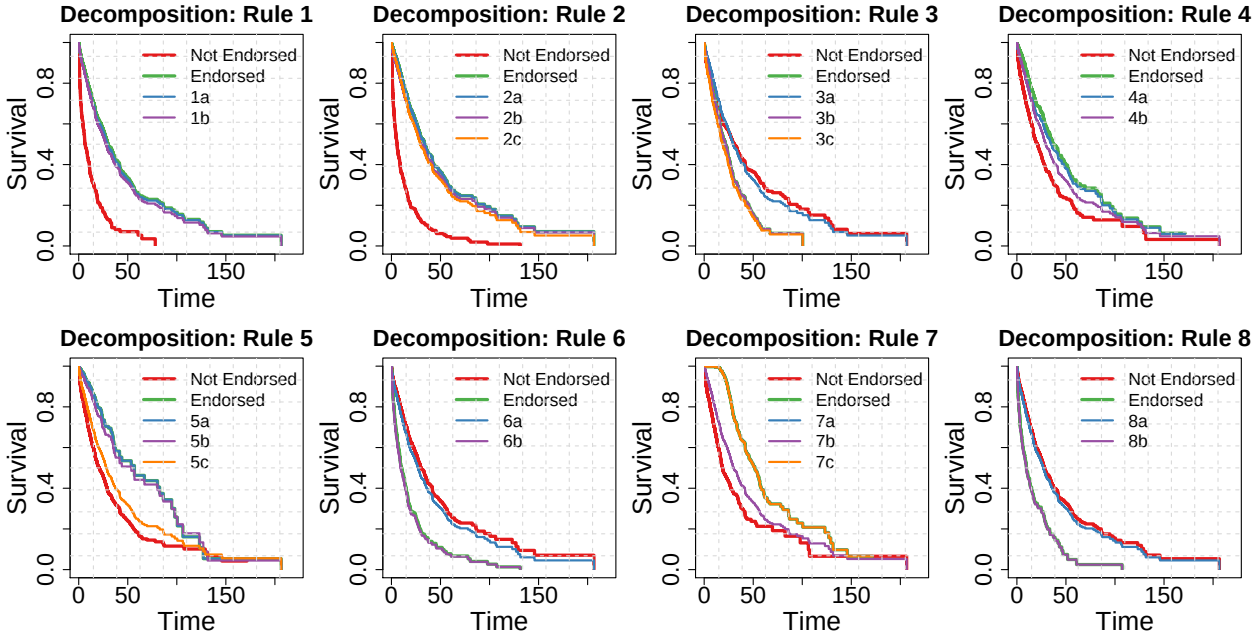


Figure 2.11: Decomposition analysis curves of rules in Table 2.4

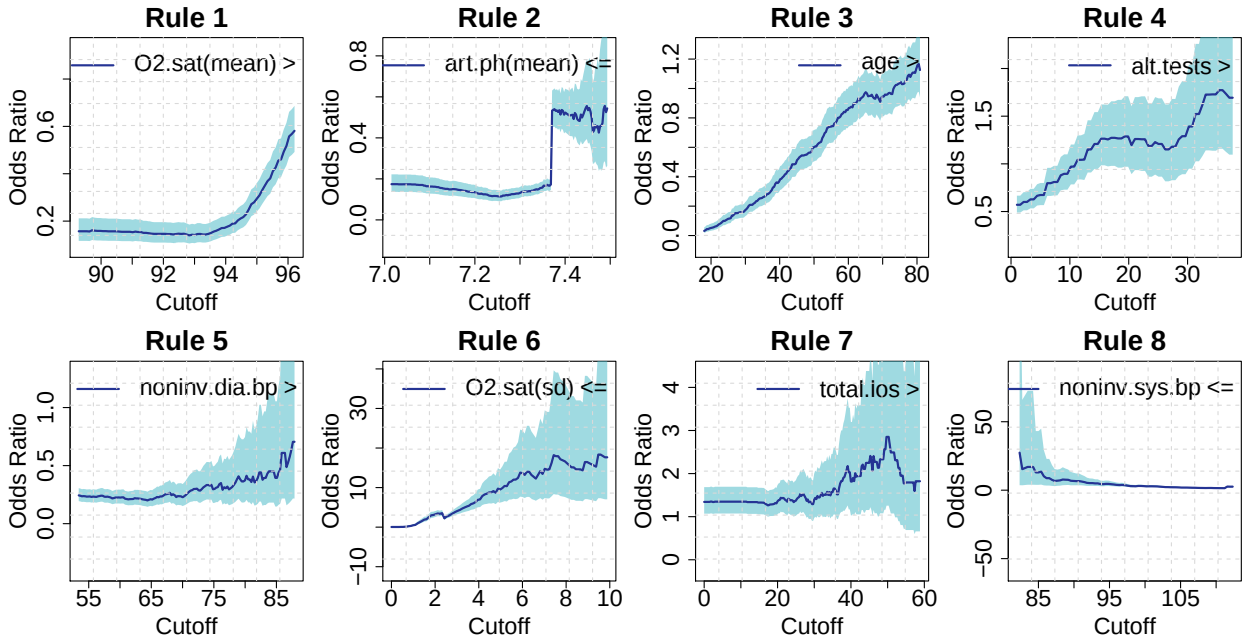


Figure 2.12: Sensitivity analysis of critical factors in rules from Table 2.4

sensitivity analysis is able to resolve such interactions and predict these complex effects. In rule 2, we see that arterial-pH below 7.2 has a slightly higher odds ratio, and therefore risk compared to when it is between 7.2 and 7.4. Any higher arterial-pH(mean) drastically increases the odds ratio implying that a high arterial-pH is a strong indicator of mortality. In rule 3, as the cutoff for age increases, the odds ratio and therefore the risk of mortality of rule endorsing observations increases steadily implying that the older population is at greater risk of mortality. Advanced age has been found to be a strong predictor of mortality among sepsis patients [90, 91]. Rules 5 and 8 show that a high diastolic or low systolic blood pressure will increase mortality risk. These insights into the effect of blood pressure are similar to those obtained by prior research in a study conducted by [83] who used blood pressure and heart rate dynamics to determine risk. Meanwhile, in rule 7, we find the longer length of stays are associated with higher risk until a stay of about 50 days, the large confidence interval of the odds ratio at stays which are any higher makes it hard to make a conclusion about the risk in this case.

2.4.5 Comparison with Cox Regression and Random Survival Forest.

We again run experiments over 100 independently sampled subsets of the sepsis data to compare the predictive performance of the SURVFIT model with the survival random forest and the Cox model. We use 4-fold cross-validation [92] to estimate the error rates of the 3 models being compared. This is done as follows: first, we divide the dataset into 4 equal and exclusive parts. Then, one of the parts is considered the test set, and the models are trained on the remaining 3 parts after which performance evaluation is done on the test set. This is done 4 times, each time considering a different part of the test set. This entire procedure is repeated 25 times for 25 different random divisions of training and the testing set to obtain the C-Index estimates for each of the different methods. The results in Fig. 2.13 show that, while the survival random forest and SURVFIT achieve comparable results, both methods significantly outperform the Cox model on this dataset.

To compare the differences in error rates by different methods, we use the paired Wilcoxon

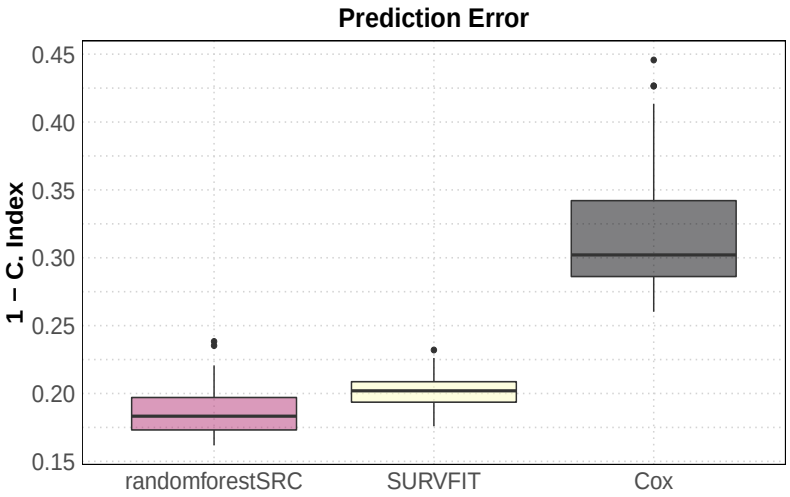


Figure 2.13: Comparison of predictive performance of Random Survival Forest and SURVFIT and Cox regression on MIMIC-III Sepsis data

rank sum test on our C-Index estimates. For each pair of the methods, we perform the following test:

Null hypothesis: $\text{C-Index}_1 = \text{C-Index}_2$
Alternative hypothesis: $\text{C-Index}_1 \neq \text{C-Index}_2$

The p-Values of each of these pairwise tests are provided in Table 2.5. It can be seen that the performance of all 3 methods are significantly different from each other. The p-Values of the tests show that the error rate of the Cox model is clearly higher than both randomforestSRC and SURVFIT, and the random survival forest method, randomforestSRC, achieves a lower error than SURVFIT. These results show that, on this dataset, SURVFIT yields greater interpretability than randomForestSRC at the cost of some prediction performance loss.

Table 2.5: p-value of pairwise Wilcoxon rank sum test on C-index obtained by each of these methods on Sepsis data

	randomForestSRC	Cox
SURVFIT	3.1e-05	2.98e-11
randomForestSRC		2.9e-11

2.5 Conclusion and Discussion

Regression models dealing with survival data such as the Cox regression model are often used as confirmative tools but are limited by their inability to discover significant interaction terms from the data unless explicitly specified. This limitation is addressed by the proposed SURVFIT method which can be used to search for significant interactions among the variables. Different from existing rule learning methods, SURVFIT extracts a doubly-sparse set of rules (i.e., which are sparse both in their cardinality as well as the cardinality of the variables involved in them) for survival data. We develop the learning formulation of SURVFIT, and further propose and evaluate fast optimization strategies. We present a rule analysis framework to analyze the extracted survival rules through statistical testing, decomposition analysis, and sensitivity analysis to draw deeper insights from them. SURVFIT could be used solely as a data analysis method that could reveal insights about the contributions and interactions of the variables. Its results could also be used to augment the Cox regression as well, i.e., with higher-order interactions. Moreover, the absence of any underlying assumptions about the data makes our model quite robust. In summary, SURVFIT provides a sparse, efficient, and highly interpretable tool that can be used to detect and explain the properties of predictive rules from survival data. We have also developed the R package, **SURVFIT**, to implement the rule learning algorithm and rule analysis framework presented in this paper. Future directions to SURVFIT may include the development of more struc-

tured solutions, such as ones with hierarchical restrictions on the variables in the rules, as well as learning rule sets such that the rule endorsement subsets are highly unique, i.e., rules are different from each other not only in terms of the variables involved but also in terms of the observations they endorse.

2.6 Software and Computational Details

R package **ranger** [57] was used to build survival random forest and **inTrees** [93] was modified by us to extract an exhaustive rule list from ranger. An R implementation of FOGLASSO based on **SLEP** [68] was used to implement the first-order method. The SOCP formulation was solved using CPLEX solver. Comprehensive codes to implement solutions of both formulations, extract survival rules, and use the proposed rule-analysis framework are available in a self-contained **SURVFIT** package downloadable from `https://github.com/hamzameer/SURVFIT`.

Chapter 3

LEARNING SEMI-PARAMETRIC BAYESIAN SURVIVAL RULE LISTS FROM HETEROGENEOUS PATIENT DATA

Survival data is often collected in medical applications from a heterogeneous population of patients. While in the past, popular survival models focused on modeling the average effect of the covariates on survival outcomes, rapidly advancing sensing and information technologies have provided opportunities to further model the heterogeneity of the population as well as the non-linearity of the survival risk. With this motivation, we propose a new semi-parametric Bayesian Survival Rule List model in this work. Our model derives a rule-based decision-making approach, while within the regime defined by each rule, survival risk is modeled via a Gaussian process latent variable model. Markov Chain Monte Carlo with a nested Laplace approximation on the Gaussian process posterior is used to search over the posterior of the rule lists efficiently. The use of ordered rule lists enables us to model heterogeneity while keeping the model complexity in check. Performance evaluations on a synthetic heterogeneous survival dataset and a real-world sepsis survival dataset demonstrate the effectiveness of our model.

3.1 Introduction

Survival analysis studies time-to-event data, and consists of important prognostic models to analyze patient morbidity and mortality in almost all medical areas. Many of these models were developed decades ago when medical data were largely collected on paper and were designed for modeling the average effect of risk factors in a population. One of the most commonly used models, the Cox proportional hazards regression [61], was built on a creative nonparametric construct of the baseline hazards function and the use of linear

formalism to characterize the relationship between the covariates and the survival outcome, i.e. hazard or risk. With increasing complexity in modern patient medical data that are now available through advances in sensor technologies, the methodological framework of the classic Cox regression model is found to be over-simplified due to its proportional hazards assumption and the imposed linearity of covariate effects. It is possible now to develop better methods for patient survival data analysis by modeling both complex survival effects as well as data heterogeneity. As these two interrelated issues are tackled in the literature as separate issues, in this paper, we propose to tackle both issues in a unified framework that builds on the strength of rule-based learning and semi-parametric Bayesian modeling for survival data. A popular way of modeling survival data is the nonparametric Kaplan-Meier estimator [94] that estimates the survival function from censored and event times but does not incorporate covariate effects. The semi-parametric Cox regression [61] can overcome this problem, but it imposes a strict proportional hazards assumption which is often not valid on real-world survival datasets. To ameliorate this limitation, the fully parametric accelerated failure time (AFT) method was proposed where a prior baseline probability distribution is given for the baseline lifetime, and covariate effects directly act on event times through a link function. Both these methods model linear effects, and can only detect interactions if explicitly specified in the model. On the other hand, the survival trees [95] and random survival forests [96] are non-parametric methods that can implicitly detect interactions and do not assume linearity. Nevertheless, being non-parametric, they are unable to incorporate prior information or quantify uncertainty. Recently, there has been work on using Bayesian semi-parametric methods for survival analysis. Gaussian process extensions to the Cox and AFT models [97, 98] have been introduced, where a latent Gaussian process is used to model one or more parameters. As GP models define distributions over functions, such models are capable of modeling non-linear effects. Moreover, since they are semi-parametric, they are capable of incorporating prior information, while the Bayesian approach enables them to quantify uncertainty. However, these models aim to model the survival data from a homogeneous population and do not model heterogeneous effects. One approach proposed

a trajectory-based collaborative modeling framework that is adaptable to model population heterogeneity for depression assessment and prognosis [99]. A prognosis-driven selective sensing method has also been proposed for faster identification of high-risk individuals in heterogeneous populations [100].

Rule-based learning is another approach that uses spatial partitioning to model heterogeneous data. Rule models have been especially important in medical and healthcare domains where interpretability is critical, e.g., rule-based machine learning models have been used for diagnosis of breast [101] and lung cancer [102], sepsis [103], diabetes [21] as well as to study depression profiles [53]. Rule-based methods have also been shown to be effective in identifying subgroups with heterogeneous risk profiles in a patient population [32]. Greedy decision tree models such as CART [33] are typical examples, but they provide a quite restrictive result, i.e., the partitioning is suboptimal. Optimal partitioning such as through integer programming [34] or Bayesian decision trees [35] is NP-hard and computationally demanding due to the exponential search space which severely limits both the depth of the tree and the number of variables that can be considered. Compared with the tree models, the rule-based methods, which work on the same principle of partitioning, have found larger flexibility and efficiency in a range of applications. As the purpose of using partitioning is to tackle patient heterogeneity, it can be used to group data into subsets with similar response characteristics that enable the subgrouping of subjects and subsequent separate modeling of each subgroup. A recent breakthrough in rule learning is Rulefit [20] for classification and regression modeling, motivated by the sparse regularization techniques, which was generalized to survival outcomes as well [25]. Rulefit generates a sparse list of predictive rules from a large set of rules mined from bootstrapped decision trees. Rules can also be flexibly used or re-organized to make better decisions, one example is the recent development of Bayesian rule lists (BRL) for classification [36] that is able to incorporate Bayesian modeling to quantify uncertainty. BRL is able to strike a balance between greedy and optimal partitioning and provides good generalizability with a significantly lower computational load.

In this paper, we propose an integrative framework that uses ordered rule lists to de-

rive a rule-based decision-making approach, while within the regime defined by each rule, survival risk is modeled via a Gaussian process latent variable model. The computational challenges are overcome by a tailored Markov Chain Monte Carlo algorithm with a nested Laplace approximation for the latent variable model to search over the posterior of the rule lists efficiently. The use of ordered rule lists enables our method to model data heterogeneity while simultaneously keeping the model complexity low and providing interpretability. Moreover, since basic GP survival models require $O(N^3)$ matrix inversion operations, the data partitioning approach may lower computational demands once a rule list is found.

3.2 Background

3.2.1 Generalized Linear Models for Survival Analysis

Different approaches aim to model the hazard function based on different premises. In the Cox proportional hazards regression [61], the hazard function is modeled as the product of a nonparametric baseline hazard, $h_0(t)$ which is a function of time and a relative hazard term, $h_R(x)$ that is a log-linear function of covariates,

$$h(t \mid x_i) = h_0(t)\exp(g(x_i)), \tag{3.1}$$

where $g(x_i) = w'x_i$. The Cox model makes an assumption that the hazards are proportional, i.e., the ratio of hazards of any two observations remains constant over time and only depends on the covariates. This restrictive assumption often does not hold on real survival datasets. Further, since the baseline hazard, $h_0(t)$ is not estimated, the Cox model can only describe how the hazards of two observations relate with each other, and not describe the hazard or survival function of a given observation directly. The accelerated failure time (AFT) model is a popular parametric approach that relaxes the proportional hazards assumption by modeling the covariate effects as directly influencing the failure time of the observations,

$$\log T_i = g(x_i) + \epsilon, \tag{3.2}$$

where ϵ is the error term with a specified distribution that determines the baseline failure density. A common modeling approach is to assign a logistic distribution to ϵ, which is equivalent to assigning a log-logistic (LL) distribution to the baseline failure time, T_0, as we can see by rewriting (3.2) as, $T_i = \exp(g(x_i))T_0$ and $T_0 = \exp(\epsilon)$. In this case, the failure time of an observation, T_i can therefore be seen as a sample from a log-logistic distribution with scale parameter, $\alpha_i = \exp(g(x_i))$ depending on covariates and a common shape parameter, β for all observations, i.e. $T_i \sim LL(\alpha_i, \beta)$, where $\alpha_i = \exp(w'x_i)$. The parameters, (w) of the Cox regression or (w, β) of the AFT model are estimated through Maximum Likelihood estimation. These two models fall under the framework of generalized linear model as the covariate effects are assumed to be linear, therefore, they cannot deal with non-linearities and interaction effects unless explicitly specified in the model terms. A popular approach to deal with some of these limitations are survival trees [95] and the bootstrap aggregation of trees, i.e., the random survival forest [96] which can implicitly deal with interactions. An added advantage of the tree-based approaches is that they are interpretable and can easily be decomposed into rules.

3.2.2 Gaussian Processes for Survival Analysis

To ameliorate the limitations of the generalized linear model framework, which can only include nonlinear effects through explicit model specification, a well-known approach has been to use Gaussian processes. Several works in literature have applied Gaussian processes to extend the standard models of survival analysis. The GP-Cox model [97] extends the Cox regression by the replacing linear effects term, $g(x)$ with a GP, f to model nonlinear covariate effects while assuming the baseline hazard to be piecewise constant. This model is extended in [98] by smoothing the piecewise constant baseline hazard with a second GP. Recently, the Gaussian Process framework was used to extend the AFT model [104] to nonlinear effects. We discuss this GP-AFT model in further detail since it is related to our work.

GP-AFT survival model

The GP extension to the log-logistic AFT model (Section 3.2.1) includes nonlinear effects by imposing a log GP prior on the scale parameter (α), instead of a log-linear relation to the variables, i.e., $\boldsymbol{\alpha} = \exp(\mathbf{f})$ where $\mathbf{f} \sim \mathcal{GP}(0, K)$. GP-AFT models the scale of the AFT distribution as dependent on the variables through the GP, while the shape parameter, β, is considered to be the same for all observations, and does not depend on covariates. However, this model carries over the AFT assumption that the shape parameter of the failure time density does not change with respect to the covariates. This is restrictive as it assumes the hazard function to be either exponential or unimodal for all observations in the data, which does not hold in heterogeneous datasets.

Work to ameliorate the limitations of Gaussian processes on heterogeneous data have focused on partitioning approaches. [105] propose Bayesian treed partitioning, with GP being fit to the terminal nodes of a Bayesian CART model, while [106] fit separate GP's at each element of a Voronoi tessellation. However, these efforts still use exclusive tree structures for modeling heterogeneity and do not address nonlinear survival models such as GP-AFT. In contrast to these fully probabilistic GP models, our work is semi-parametric which enables us to achieve a balance between computational demand and performance. Our proposed work also addresses the limitation of the GP-AFT survival to heterogeneity by relaxing the common shape assumption and varying the shape parameter with respect to covariates, i.e., our rule list approach allows us to learn failure time densities with varying shapes for different partitions, as well as identify the covariates that cause heterogeneity.

3.3 GPSRL

Let $\mathcal{R}$ be the pre-mined rule set containing a total of K rules. We generate $\mathcal{R}$ by extracting rules of various cardinalities from trees in a random survival forest. The cardinality of a rule is defined as the number of interacting covariates in the rule (ex. $1, 2, ..$ etc.). Rules that are endorsed by at least a given minimum number of observations in the dataset are selected, i.e.,

rules that apply to very few observations are filtered out. Our goal is to tackle heterogeneity by building an *ordered* rule list, d which is a subset of $\mathcal{R}$ of size m where $m \ll K$. Priors on the number of rules in the list, m, and the number of covariates interacting in each rule ensure that the rule list is sufficiently sparse and complex. We utilize an MCMC scheme similar to that in BRL [36] to obtain a posterior over the ordered rule list given data; however, the objective in BRL was multivariate classification, while our goal is survival analysis with Gaussian processes. In what follows, we describe our predictive model, Gaussian process survival rule lists (GPSRL), and inference procedure to learn the model.

3.3.1 Formulation of Gaussian Process Survival Rule Lists

An ordered rule list, d with m rules will divide the dataset into $m+1$ non-overlapping partitions as follows: each observation in the dataset which endorses at least one of the m rules belongs to the partition associated with the *first* rule in the ordered list, d, that the observation endorses. Observations not endorsing any of the m rules will belong to the $m+1$-th partition. Thus, an ordered rule list of length m will divide the data into $m+1$ exclusive partitions. For each of the partitions determined by d, we fit a log-logistic (LL) Gaussian process AFT model (3.2.2) with its scale parameter ($\boldsymbol{\alpha}$) dependent on covariates and modeled via a Gaussian process, and a shared shape (β) parameter with a log uniform prior. The same priors are adopted across partitions to control the model complexity. An illustration of our Bayesian GPSRL model is given in (3.3).

$$\begin{aligned} &\textit{if } r_1 \textit{ then } t_1 \sim LL(\boldsymbol{\alpha_1}, \beta_1) \\ &\textit{else if } r_2 \textit{ then } t_2 \sim LL(\boldsymbol{\alpha_2}, \beta_2) \\ &\qquad . \\ &\qquad . \\ &\textit{else if } r_m \textit{ then } t_m \sim LL(\boldsymbol{\alpha_m}, \beta_m) \\ &\textit{else } t_0 \sim LL(\boldsymbol{\alpha_0}, \beta_0) \end{aligned} \tag{3.3}$$

The priors on the parameters of the rule list and survival models on each partition ($j \in \{0, \cdots m\}$) are:

$$
\begin{aligned}
& m \sim TP(\lambda, 0, |R|) && \boldsymbol{\alpha_j} = \exp\big(f_j(X_j)\big) \\
& \log\, \beta_j \sim U(0, s) && f_j(X_j) \sim \mathcal{GP}(0, K_j) \\
& s \sim IG(a, b) && K_j(x, x) = \sigma_j^2 \exp\big(-\frac{|x - x'|^2}{2l_j^2}\big) \\
& \sigma_j \sim IG(a_\sigma, b_\sigma) && l_j \sim IG(a_l, b_l)
\end{aligned}
\tag{3.4}
$$

Here, LL, U, TP, IG, N are the Log-logistic, Uniform, Truncated-Poisson, Inverse-Gamma and Normal distributions, respectively. Our model seeks to combine the interpretability of rule-based models with the modeling flexibility of Gaussian process survival models. Given the pre-mined set of rules $\mathcal{R}$, we seek to obtain the posterior distribution of the ordered Bayesian rule list, d, and the associated posterior distributions of parameters of the Gaussian survival processes that model the survival response at each of the corresponding partitions defined by d.

3.3.2 Bayesian inference

Given covariate data, X, and survival response, y, Our goal is to obtain the posterior distribution of the ordered Bayesian rule list. The posterior probability density of the rule list, d is proportional to the product of data likelihood and prior probability:

$$p(d \mid X, y) \propto p(y \mid X, d)p(d). \tag{3.5}$$

Prior probability

Similar to the prior probability of the BRL [36] model, the prior probability of the GPSRL rule list is defined hierarchically as,

$$p(d) = p(m \mid \lambda) \prod_{j=1}^{m} p(c_j \mid c_1 \cdots c_{j-1}, \eta) p(r_j \mid r_1, \cdots r_{j-1}, c_j). \tag{3.6}$$

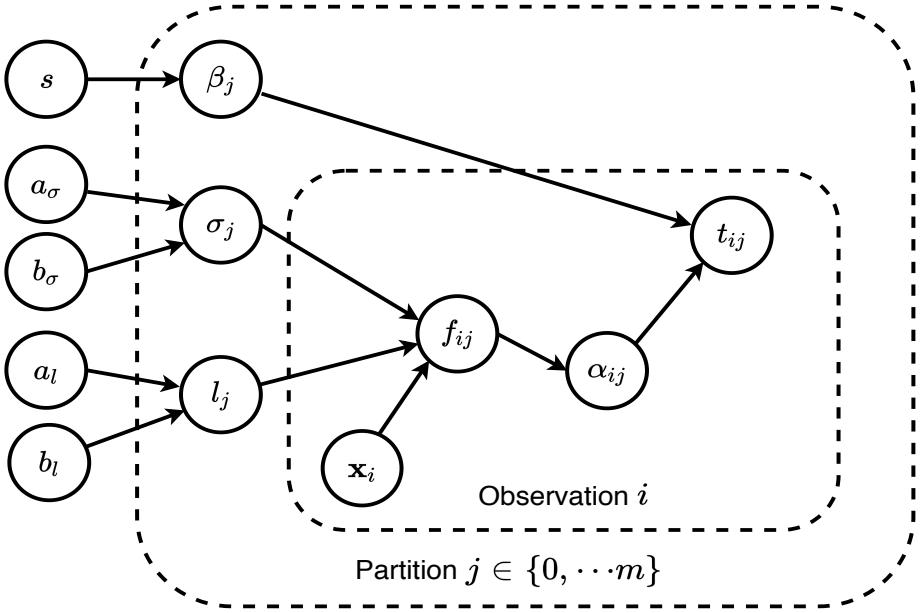


Figure 3.1: Graphical representation of GPSRL

Here m is the number of rules in the list, and c_j denotes the cardinality of rule r_j. Truncated-Poisson (TP) priors are selected for both m and each of $c_j \mid c_1 \cdots c_{j-1}$. The TP prior on m has a given mean value λ and is truncated on the total number of rules that are available, K. It may be written as follows.

$$p(m \mid \mathcal{R}, \lambda) = \frac{(\lambda^m / m!)}{\sum_{j=1}^{K} (\lambda^j / j!)}, \quad m = 0, 1, \ldots K \tag{3.7}$$

The TP prior on $c_j \mid c_1 \cdots c_{j-1}$ has a mean η and is truncated to only include the cardinalities of rules that are currently available. $\mathcal{R}_j = \mathcal{R} \setminus \{r_1, r_2 \ldots r_{j-1}\}$ are the rules available after sampling $j-1$ rules. The probability of selecting the cardinality, c_j may be written as:

$$p(c_j \mid c_1 \cdots c_{j-1}, \eta) = \frac{\eta^{c_j} / c_j!}{\sum_{c \in C_j} \eta^c / c!}, \tag{3.8}$$

where C_j is the set of cardinalities of rules in $\mathcal{R}_j$. Once c_j is sampled, a uniform probability is chosen over all the available rules with cardinality c_j to sample the j-th rule in d, r_j. That is,

$$p(r_j \mid r_1, \cdots r_{j-1}, c_j) = \frac{1}{|\{r_i \mid r_i \in \mathcal{R}_j, c_i = c_j\}|} \tag{3.9}$$

The first term in the prior (3.6) is the probability of obtaining a rule list with m rules given the mean number of rules λ. In the subsequent product over the rules $\{r_1 \cdots r_m\}$ in d, each term denotes the probability of obtaining a rule r_j with a cardinality of c_j given a mean cardinality η multiplied by the probability of choosing rule r_j from all the available rules with this cardinality.

Likelihood

We fit a log-logistic (LL) Gaussian process AFT model (3.2.2) at each of the partitions defined by the rule list, d. Here, the data likelihood (**??**) of partition $j \in \{1, 2, \ldots m+1\}$ given response $y = (t, \boldsymbol{\delta})$, covariate data, X, and parameters $(\boldsymbol{\alpha}, \beta)$ is as follows:

$$p(y \mid \boldsymbol{\alpha}, \beta) = \prod_{i=1}^{M:\boldsymbol{\delta}=1} \frac{(\frac{\boldsymbol{\alpha}_i}{\beta})(\frac{y_i}{\boldsymbol{\alpha}_i})^{\beta-1}}{(1+\frac{y_i}{\boldsymbol{\alpha}_i}\beta)^2} \prod_{j=1}^{N:\boldsymbol{\delta_j}=0} \frac{1}{1+(\frac{y_i}{\boldsymbol{\alpha}_i})^{\beta}}. \tag{3.10}$$

The scale parameter, $\boldsymbol{\alpha} = \exp(\mathbf{f})$ is defined as an exponential of a Gaussian process (GP) with prior $\mathbf{f} \mid X \sim \mathcal{GP}(0, K_X)$, while the shape parameter has a log uniform prior, $\log\, \beta \sim U(0, s)$. Since this likelihood (3.10) is not Gaussian or conjugate-Gaussian, the posterior density of $\mathbf{f}$, i.e., $p(f \mid X, y)$, and consequently the marginal data likelihood, $p(y \mid X) = \int p(y \mid f)p(\times f \mid X)df$ is not analytically tractable and must be approximated. A popular method to approximate the posterior in case of non-Gaussian likelihood with latent Gaussian processes such as those arising in survival analysis is the Laplace approximation, which obtains a Gaussian distribution approximation to the posterior density of the GP around the mode of the true distribution. The Laplace approximation [107] obtains a Gaussian density, q, which approximates the true non-Gaussian posterior density of the GP, i.e., the approximate posterior density, $q(f) \approx p(f \mid y, X)$ given by,

$$q(f) = \mathcal{N}(f \mid \hat{f}, A^{-1}) \tag{3.11}$$

where $\hat{f} = \text{argmax}_f p(f \mid X, y)$ is the mode of the posterior and $A = -\nabla\nabla \log\, p(f \mid X, y)$ is the Hessian of the negative log posterior at the mode. This approximate density can be

used in lieu of the true GP posterior to calculate an approximated marginal likelihood, $p(\times y \mid X) \approx p_L(y \mid X) = \int p(y \mid f)q(f)df$. The total approximated marginal likelihood of the data can be written as:

$$p_L(y \mid \mathbf{X}, d) = \prod_{j=0}^{m} p_L(y_j \mid \mathbf{X}_j), \tag{3.12}$$

where y_j and X_j is the response and covariate data, respectively, belonging to each of the $m+1$ partitions (indexed by j) defined by our rule list.

MCMC Sampling

We use Metropolis-Hastings sampling to infer the posterior distribution of the rule list, $p(d \mid X, y)$. The sampling sequence of rule lists starts with an initial random list, d^0 sampled from the prior, $p(d)$. After initialization, the sequence proceeds as follows: At step t in the sequence, with a rule list d^t of length m^t, a proposal distribution Q is used to propose the next list in the sequence $d^{t+1} \sim Q(d^t)$. The new rule list is generated through one of three equally likely operations: i) adding a rule to the bottom of d^t, ii) removing a randomly selected rule from d^t, iii) moving a randomly selected rule to a different position in d^t. The proposal distribution denotes the probability of sampling d^{t+1} from d^t,

$$Q(d^{t+1} \mid d^t, \mathcal{R}) = \begin{cases} \frac{1}{(K-m^t)(m^t+1)} & \text{if a rule is added} \\ \frac{1}{m^t} & \text{if a rule is removed} \\ \frac{1}{m^t(m^t-1)} & \text{if a rule is moved.} \end{cases} \tag{3.13}$$

The proposed sequence, d^{t+1} is then accepted with an acceptance probability, $\pi(d^{t+1} \mid d^t)$, defined as follows:

$$\pi(d^{t+1} \mid d^t) = \min\Big\{ \frac{Q(d^{t+1}, d^t)}{Q(d^t, d^{t+1})} \frac{p(\mathbf{y} \mid \mathbf{X}, d^{t+1})p(d^{t+1})}{p(\mathbf{y} \mid \mathbf{X}, d^t)p(d^t)}, 1 \Big\}. \tag{3.14}$$

Here $p(y \mid X, d^t)$ and $p(y \mid X, d^{t+1})$ are marginals that may be evaluated approximately as shown in Section 3.3.2. For a sufficiently long chain, the sequence will sample rule lists from the posterior density of the Bayesian rule list. Gaussian approximates to the marginal

likelihood have been proposed in literature to increase the speed of the MCMC algorithm when the likelihood evaluation is costly [108], where it was shown that using an approximate likelihood may take more MCMC steps to reach convergence though the total time of convergence reduces due to faster sampling enabled by the approximations. However, using approximate likelihood evaluations does not theoretically guarantee convergence of the MCMC algorithm though the algorithm will push the sequence towards an area with a high approximate posterior.

3.3.3 Predictive Inference

Given the posterior distribution of the rule lists obtained from the MCMC sequence, $p(d \mid \times X, y)$, we can obtain a point estimate of the rule list, d, and the model defined by the Laplace-posterior approximations of the latent GP, $q_j(f)$ and shape parameter of the log-likelihood distribution, β_j at each of the partitions. The predictive density (under the Laplace approximation) of a new observation (y^*, x^*) which falls in, say, the j-th partition as defined by d can be evaluated as follows: first, the distribution of the latent GP at the new observation (see [107] for derivation) under the Laplace approximation is computed. Since,

$$\begin{bmatrix} f^* \\ f \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X) & K(X, x^*) \\ K(x^*, X) & K(x^*) \end{bmatrix}\right) \tag{3.15}$$

and therefore the conditional density is

$$f^* \mid x^*, X, f = \mathcal{N}\Big(K(x^*, X)K(X)^{-1}f, K(x^*) - K(x^*, X)K(X)^{-1}K(X, x^*)\Big). \tag{3.16}$$

The posterior predictive density of the GP under the Laplace approximation, $q_j(f^* \mid x^*, \times X, y)$, can then be evaluated as,

$$\begin{aligned} p_j(f^* \mid X, y, f) &= \int p_j(f^* \mid x^*, X, f)p(f \mid X, y)df \\ &\approx \int p(f^* \mid x^*, X, f)q_j(f \mid X, y)df \\ &= q_j(f^* \mid x^*, X, y) \end{aligned} \tag{3.17}$$

and is therefore a Gaussian distribution that is analytically tractable. Then, the predictive density of the observation (y^*, x^*) under the Laplace approximation is given by the integral,

$$\begin{aligned} p(y^* \mid x^*, X, y) &= \int p(y^* \mid \alpha^*, \beta_j) p_j(f^* \mid x^*, X, y) df^* \\ &\approx \int p(y^* \mid \alpha^*, \beta_j) q_j(f^* \mid x^*, X, y) df^* \end{aligned} \tag{3.18}$$

Since the first term of this integral is the log-logistic data likelihood (3.10), and is not conjugate-Gaussian, the integral is not analytically tractable and must be calculated using numerical methods.

3.4 Numerical Experiments

Performance validation and comparison with several existing GP-based survival models such as GP-AFT survival Gaussian process model with Laplace approximation [109] SGP(L), survival GP model with variational approximation [110] SGP(V), and a recent work that proposed to model both shape and scale parameters with GP's, chained Gaussian process model [111], CHGP. Experiments are performed on a synthetic heterogeneous survival dataset and a real-world survival dataset of sepsis patients from the MIMIC-III (Medical Information Mart for Intensive Care) [76] database. Performance is evaluated using the negative log predictive density (NLPD) (3.18) and concordance index (C-INDEX) [75]. NLPD is calculated from the predictive density as shown in Section 3.3.3,

$$\text{NLPD}(y^*, X^*) = \frac{\sum_{i=1}^{N} p(y_i^* \mid x_i^*, X, y)}{N}. \tag{3.19}$$

C-INDEX is the measure of a model's ability to rank survival times. It estimates the probability that in a randomly selected pair of test observations, the one with the lower response time has the lower predictive response time. In our experiments, for each observation, we take the average C-INDEX calculated over 100 predicted times sampled from the predictive log-logistic distributions in the obtained GPSRL rule lists (3.3). NLPD is lower in a superior model while C-INDEX is higher. We use the GPy [112] software to train the GP models used in these experiments.

3.4.1 Synthetic Data

We simulated a heterogeneous survival dataset, $D = (y, \delta, X)$, consisting of $N = 1000$ observations with $P = 4$ variables. The covariates of each observation $x \in X$ are generated by sampling from uniform distribution ($x_i \sim U(0,1)\, \forall i \in 1 : P$). Event times for all observations, t, are simulated by sampling from a log-logistic distribution, $t_i \sim LL(\alpha, \beta)$ with the scale α, and shape β parameters generated as follows:

$$\alpha(x) = I_1\alpha_1(x) + I_2\alpha_2(x) + I_3\alpha_3(x),$$
$$\beta(x) = I_1\beta_1(x) + I_2\beta_2(x) + I_3\beta_3(x),$$

where I_1, I_2, I_3 are indicator functions to denote certain conditions on the covariate data being satisfied and α_i, β_i are different complex functions of the covariates of the following form:

$$\alpha_i(x) = a_1 \exp\Big(a_2\big(\sum_{k=1}^{2} \exp(a_3(x[k] - a_4)^2)\big) + \sum_{k=3}^{4} \sin(\pi x[k]^2)\Big),$$

$$\beta_i(x) = b_1 \exp\Big(\sum_{k=1}^{2} \sin(2\pi x[k]^2) + \sum_{k=3}^{4} \cos(2\pi x[k]^2)\Big),$$

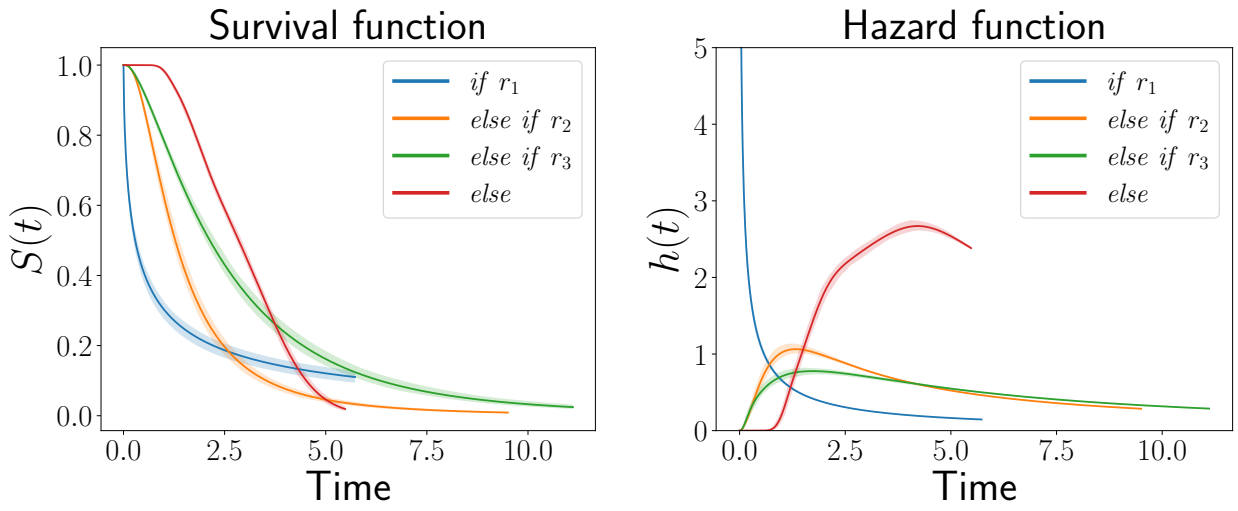
with varying values of a_1, a_2, a_3, a_4 and b_1 for each $i \in \{1, 2, 3\}$. We assume that 35% of the simulated data is censored ($\boldsymbol{\delta}_i = 0$). To account for this, the simulated event times, t of a random subset consisting of 35% of the data are multiplied with a uniform random variable to simulate the response times, y, i.e., $y_i = \rho_i t_i$ if $\boldsymbol{\delta}_i = 0$ else $y_i = t_i$ where $\rho_i \sim U(0,1)$. Performance comparison is carried out by evaluating model performance on a testing dataset consisting of 250 observations that was simulated in a similar manner as the training data. A large set of rules are mined from a survival random forest and rules that are followed by at least 10% of the data are selected to generate the initial rules, $\mathcal{R}$.

Results

We used hyperparameter values of $\lambda = 3$ for the mean length of the rule list, and $\eta = 2$ for the mean number of variables in each of the rules. The MCMC chain was simulated until

Table 3.1: Estimate of ordered rule list d from the posterior

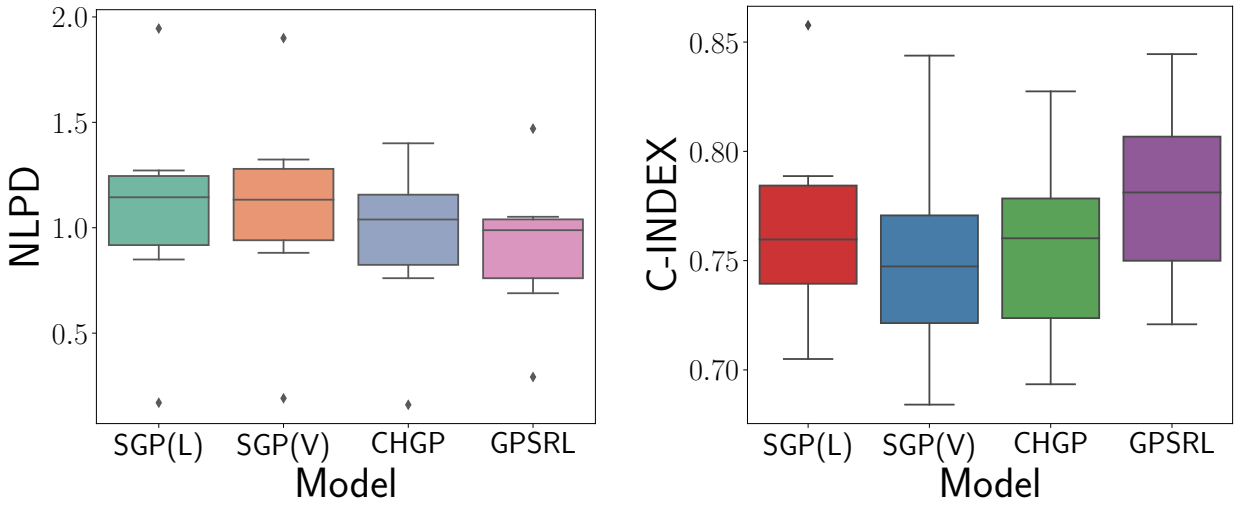
Rules	
r_1	$x_3 \leq 0.259$
r_2	$x_4 > 0.596$ & $x_3 > 0.196$
r_3	$x_4 \leq 0.677$ & $x_4 > 0.192$



(a) (b)

Figure 3.2: (a) Mean survival function and (b) Mean hazard function of GPSRL model on each of the partitions defined by rule list in Table 3.1

convergence, which took approximately 4000 iterations. An example of the rule list obtained from the posterior in one of the MCMC chains is shown in Table 3.1. Fig. 3.2a shows the mean survival function learned at each of the four data partitions defined by the rule list, and Fig. 3.2b shows the mean hazard function. It is noted that the hazard function learned by GPSRL is multimodal, i.e., in the first partition, the hazard is exponential, signifying that the shape parameter, $\beta <= 1$ while in the other partitions, it is unimodal meaning $\beta > 1$. The standard GP-AFT models discussed in Section 3.2 assume the same value of β for all observations and hence learn either an exponential or a unimodal hazard but not both, while the partitioning approach of GPSRL allows us to model both unimodal and multimodal hazards, which is a typical aspect of heterogeneous medical datasets.



(a) (b)

Figure 3.3: (a) NLPD and (b) C-INDEX comparison over 10 cross-validation folds with 250 replicates for different survival GP models trained on synthetic data

The performance comparison results in Fig. 3.3 show that GPSRL model outperforms the other models on the synthetic dataset. The box plot in Fig. 3.3a shows the NLPD obtained by each model on the testing data over 10 folds of the testing data. We observe that

GPSRL achieves the lowest values of NLPD as compared to other models, demonstrating its effectiveness. The other survival model that provides a way to model heterogeneity, chained survival Gaussian process performs second best. The GP-AFT models with Laplace approximation, SGP-L and with sparse GP variational approximation, SGP-V achieve comparable performance. In Fig. 3.3b, the comparison of C-INDEX for different models on the 10 testing folds is shown. Once again, GPSRL outperforms other models and has the highest mean C-INDEX. The C-INDEX follows the same trend as NLPD, however, SGP(V) achieves the lowest C-INDEX.

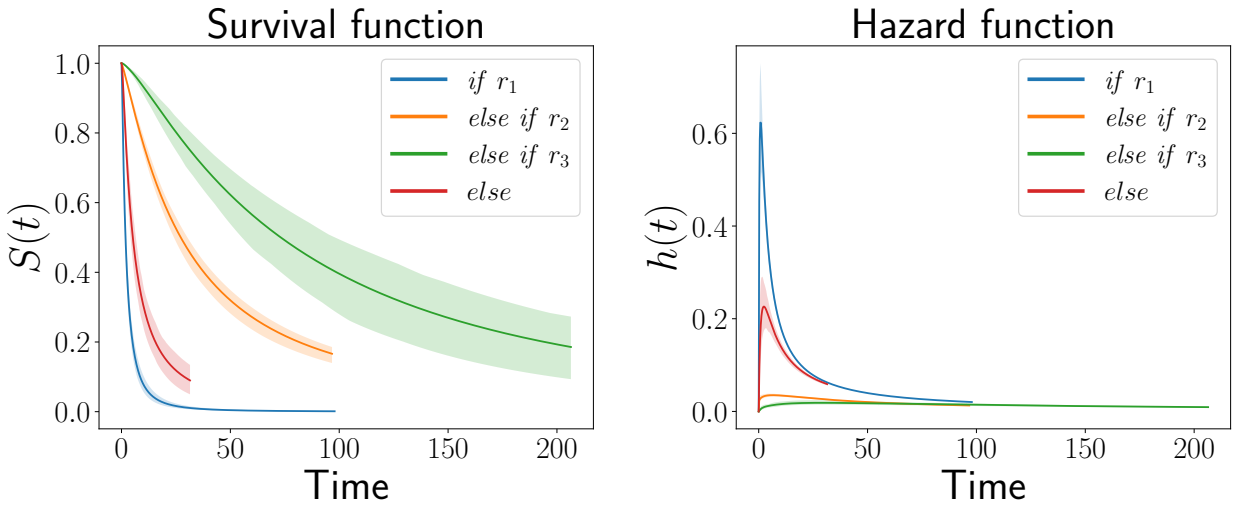
3.4.2 Sepsis Data

MIMIC-III is a comprehensive database comprising anonymized information relating to patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA between 2001 and 2012. The data consists of over 53,000 adult ICU admissions during this time period. In this paper, we utilize a subset of inpatient admissions that were diagnosed with sepsis conditions. We consider 9 variables consisting of patient characteristics and physiological measurements, which are age, heart rate, diastolic and systolic blood pressure, saturated oxygen, arterial-pH etc. We choose a subset of 1200 observations for training the model and a testing dataset of 400 observations.

Table 3.2: Estimate of ordered rule list d from the posterior

Rules	
r_1	artpH-(mean) <= 7.249
r_2	O2sat-(sd) <= 4.66 & diaBP-(mean) <= 61.26
r_3	O2sat-(sd) <= 4.8

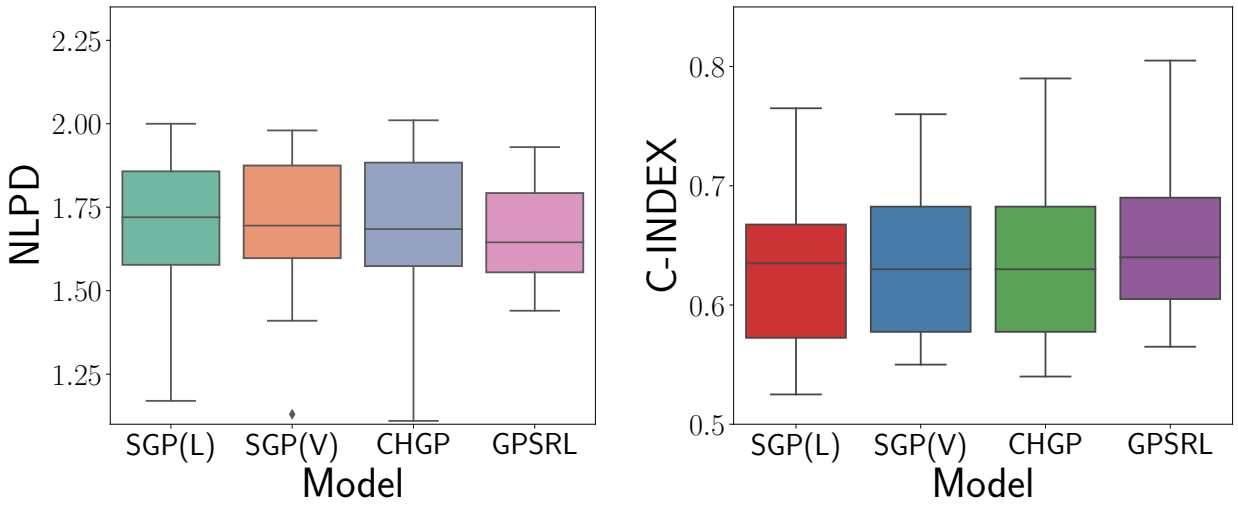
Results



(a) (b)

Figure 3.4: (a) Mean survival function and (b) Mean hazard function of GPSRL model on each of the partitions defined by rule list in Table 3.2

We used hyperparameter values of $\lambda = 3$ for the mean length of the rule list, and $\eta = 2$ for the mean number of variables in each of the rules. The MCMC chain was simulated until convergence, which took approximately 4700 iterations. An example of the rule list obtained from the posterior in one of the MCMC chains is shown in Table 3.2. Fig. 3.4a shows the mean survival function learnt at each of the three data partitions defined by the rule list, and Fig. 3.4b shows the mean hazard function. The performance comparison results in Fig. 3.5 show that on this dataset, the performance of all the models is comparable though a small gain is achieved by using GPSRL. The box plot in Fig. 3.5a shows the NLPD achieved by each model on the testing data over 10 folds of 40 observations each. GPSRL does achieve a slightly lower median NLPD as compared to other models, though standard GP models are satisfactory on this dataset. The same applies to Fig. 3.5b which compares the C-INDEX



(a) (b)

Figure 3.5: a) NLPD and (b) C-INDEX comparison over 10 cross-validation folds with 400 replicates for different survival GP models trained on sepsis data.

achieved by the models on the 10 testing folds. Once again, the values obtained are nearly equal.

Table 3.3: Summary of NLPD comparison of different models

Data	SGP(L)	SGP(V)	CHGP	GPSRL
Synthetic	1.08 ± 0.2	1.09 ± 0.2	0.93 ± 0.1	0.9 ± 0.1
Sepsis	1.76 ± 0.2	1.75 ± 0.2	1.73 ± 0.2	1.67 ± 0.2

A summary of the mean NLPD obtained via cross-validation by the various survival GP models on both the synthetic and survival test datasets is provided in Table 3.3, and a summary of the mean C-INDEX is provided in Table 3.4. As can be seen, our model achieves clearly better performance in both average NLPD and average C-INDEX on the synthetic

Table 3.4: Summary of C-INDEX comparison of different models

Data	SGP(L)	SGP(V)	CHGP	GPSRL
Synthetic	0.77 ± 0.05	0.75 ± 0.05	0.75 ± 0.05	0.78 ± 0.04
Sepsis	0.61 ± 0.07	0.62 ± 0.06	0.62 ± 0.07	0.63 ± 0.07

heterogeneous dataset while all the GP models have more or less similar performance on the sepsis data.

3.5 Conclusion

In this paper, we propose a novel and effective method to model heterogeneity in survival data analysis. Our model, 'Gaussian Process Survival Rule Lists (GPSRL)', utilizes a semi-parametric Bayesian framework to partition the data into subsets with different survival characteristics. This allows us to address some of the limitations of standard survival Gaussian process models and also provides a degree of interpretability. Experimental results on the synthetic dataset and the MIMIC sepsis survival dataset demonstrate the efficacy of our model by outperforming existing survival GP models. In future work, exploring speedups to GPSRL through either screening for bad proposals of rule lists or computationally efficient approximations, such as stochastic approximations to the latent marginal can improve the model further.

Chapter 4

A RULE-BASED EXPLORATORY ANALYSIS FOR DISCOVERY OF MULTIMODAL BIOMARKERS OF ADHD USING EYE MOVEMENT AND EEG DATA

Developing biomarkers for a complex neurodevelopmental disorder such as attention deficit hyperactivity disorder (ADHD) is a challenging task since it is a multifactorial and multi-faceted condition. Researchers have been employing different sensing modalities to acquire measurements of the condition, however, there has been a lack of approaches that can adequately combine the multimodal data and detect interactions among the modalities. To demonstrate the concept and benefit of multimodal biomarker discovery, we conducted a multimodal data collection targeting the ADHD condition and demonstrated how a rule-based exploratory analysis approach could be used to analyze the data. To the best of our knowledge, our work is the first attempt to explore and identify interesting interactions among two modalities of data, eye movement data and the EEG signal, for multimodal biomarker discovery for ADHD. The detection of these interactions would help us better understand the condition and develop better prediction models and intervention strategies.

4.1 Introduction

Attention Deficit/Hyperactivity Disorder (ADHD) is a common neurodevelopmental disorder, which is usually first diagnosed in childhood but may last well into adulthood [52, 113, 114]. ADHD is characterized by an ongoing pattern of inattention or impulsivity and can significantly interfere in the functioning and development of individuals [52]. While clinical diagnosis of ADHD is based on the DSM-IV diagnostic criteria [115], researchers have been trying to understand the underlying pathology, and to develop ADHD prediction models

based on statistical and machine learning methods. As opposed to the subjective and self-reported observations of behavior required in the DSM criteria, statistical methods could enable the detection of objective biomarkers that are predictive of ADHD. For example, brain imaging methods have been explored as diagnostic aids [116, 50, 117] based on MRI images and resting-state fMRI scans. Facial expression and head motion analysis via RGB-D depth-sensing camera and based on dynamic deep learning [118] has also been explored as a potential diagnostic aid. Further, models based on continuous performance test data [119] that include presentation of visual or auditory target and non-target stimuli aimed at measuring inhibition and impulsivity [120] were found to enhance diagnostic assessment and could be used as supplementary measures in ADHD diagnosis.

Among these efforts, an important area is the use of electroencephalography (EEG) in detecting ADHD patterns [43, 44, 45, 46]. EEG data has been used to describe and quantify the underlying neuro-physiology of ADHD as well as in many diagnostic applications [43]. For example, EEG studies have discovered that elevated theta-beta ratio is a commonly observed trait in participants with ADHD [44, 121]. Further, children with ADHD have been found to have reduced power in alpha and beta bands and increased power in the low EEG frequency bands (i.e., delta and theta bands) in comparison to children without ADHD [122].

More recently, eye movement data has also been used in developing prediction systems for ADHD [47, 48, 49], since eye movement patterns may reveal attention and working memory deficits that are prevalent in individuals suffering from ADHD. Recent work integrating an eye tracker with continuous performance tests (CPTs) was proposed to be a feasible way of enhancing diagnostic precision [47], and a combination of eye movement fixation and saccade features have been shown to be predictors of ADHD diagnosis in adults [123, 124]. Further, a decision tree model [125] that uses eye movements and positions of different gaze event types such as fixations, saccades, gaze positions, and pupil diameters has been found to have good predictive accuracy.

So far, there has not been a study that jointly explored eye movement data and EEG

signals for ADHD detection. ADHD has a complex pathology, and one single modality is not sufficient to characterize it completely. Multimodal biomarker analysis has proven to have superior predictive accuracy in the modeling of Alzheimer's disease [126]. Therefore, using multimodal biomarkers may not only enhance the prediction power of the diagnostic model of ADHD, but also deepen our understanding of the condition and further yield better design of intervention strategies. Some existing works using multimodality to detect ADHD focused on the fusion of fMRI and EEG data [50, 43] but not EEG and eye movement data. We note that fMRI is restricted to be used in constrained environments where participants have to be immobile in a supine position [127]. This limits the usage of fMRI for complex tasks in everyday situations [128], which is critical to diagnose ADHD [127]. In prior works, EEG and eye movement data were combined for the purpose of better data preprocessing, e.g., [129, 130] have proposed methods for correcting the artifacts generated by eye movement in EEG signal using simultaneous eye tracking information, but a joint study considering multimodal effects has not yet been undertaken. The main challenge for detecting multimodal biomarkers for ADHD is the lack of data where multimodal data should be simultaneously collected, and also a lack of computationally efficient models that can efficiently sweep through the enormous search space of potential multimodal interaction patterns. In this work, we seek to address these challenges by leveraging on recent developments in machine learning, and particularly, the development of rule learning methods [131, 132] that yield high-quality predictive rules and conduct an exploratory analysis of a multimodal dataset of ADHD patients we have recently collected for this endeavor.

This paper [133] is structured in the following manner. Section 4.2 describes the data and the study population. Section 4.3 presents the rule learning and analysis framework that we use to analyze the data. Section 4.4 presents the prediction performances of different models, the discovered rules and their decomposition analysis. Section 4.5 gives an in-depth analysis and interpretation of some rules and their implications on better ADHD prediction and intervention design. Section 4.6 concludes the study.

4.2 Data

We collected the eye movement and EEG data from twenty-two undergraduate and graduate students enrolled at the University of Washington who were recruited to carry out the task and collect the eye movement and EEG data. The sample consisted of 9 male and 13 female students with a mean age of 20.4 years. The students had no prior knowledge of the task, and the study was approved by the University Institutional Review Board (IRB). Three of the students were diagnosed with ADHD. Prior to the main experiment, participants were asked to answer two questions: Have you been diagnosed with ADHD by a medical professional? and have you been diagnosed with a different disorder, such as autism or ADD, by a medical professional? Three students responded 'yes' to question 1. Out of 3 students who responded that they had been diagnosed with ADHD by a medical professional, one participant responded with autism and the other responded with anxiety to question 2.

The study used NASA's Multi-Attribute Task Battery (MATB-II) to simulate a multi-tasking environment for participants in the experiments [134], which has been widely used to evaluate operators' multitasking performance and workload [134]. The participants needed to simultaneously perform system monitoring, communications, and resource management tasks, which were analogous to tasks that aircraft pilots perform in flight. During the experiments, MATB-II displayed task stimuli randomly, and participants were required to react to each stimulus by clicking on the screen with a mouse as soon as possible. Reaction time, and correct and incorrect responses were recorded. During the experiments, the participants first familiarize themselves with the MATB-II tasks in a five-minute training session. Then, the participants performed four eight-minute experiment sessions. We randomized the order of sessions for all participants to eliminate the order effect. Table 4.1 summarizes the basic statistics of the participants, and a screenshot of the task interface and the participant's eye movement is shown in Figure 4.1. The participants had no prior knowledge of the task, and the study was approved by the University Institutional Review Board. All participants signed a written consent form prior to the experiments. Since each participant had done 4

different tasks, in total we have 88 samples, while among them 12 samples are observed on ADHD subjects.

Table 4.1: Basic statistics of the study participants.

	Without ADHD	With ADHD
Number of participants	19	3
Mean (standard deviation) age	20.42 (2.04)	21.33 (0.58)
Gender		
Male	6 (31.6%)	1 (33.3%)
Female	13 (68.4%)	1 (33.3%)
Not identified	0	1 (33.3%)
Race and ethnicity		
Latino or Hispanic	3 (15.8%)	0
Asian	4 (21.1%)	2 (66.7%)
White/Caucasian	9 (47.7%)	1 (33.3%)
More than one	3 (15.8%)	0

4.2.1 Eye movement data

We used a Tobii Pro X3-120 eye tracker (120 Hz sampling rate, 0.6 degrees of gaze accuracy) to track and record participants' eye movements. The eye tracker was mounted below the screen of a 15-inch laptop (1920 px $\times$ 1080 px resolution) during the experiments. The average viewing distance between participants and the laptop screen was 40 cm. We used the Tobii Studio software to calibrate participants' eye movements before each experiment session and output the eye movement data for analysis. The eye tracker along with the Tobii Studio software collected participants' gaze and pupil information during the experiments and classified the gazes as fixations and saccades. Figure 4.1 visualizes part of a participant's

gaze on the screen. Each node represents a fixation and each edge between nodes represents a saccade. The numbers on the circles indicate the sequence of the fixations.

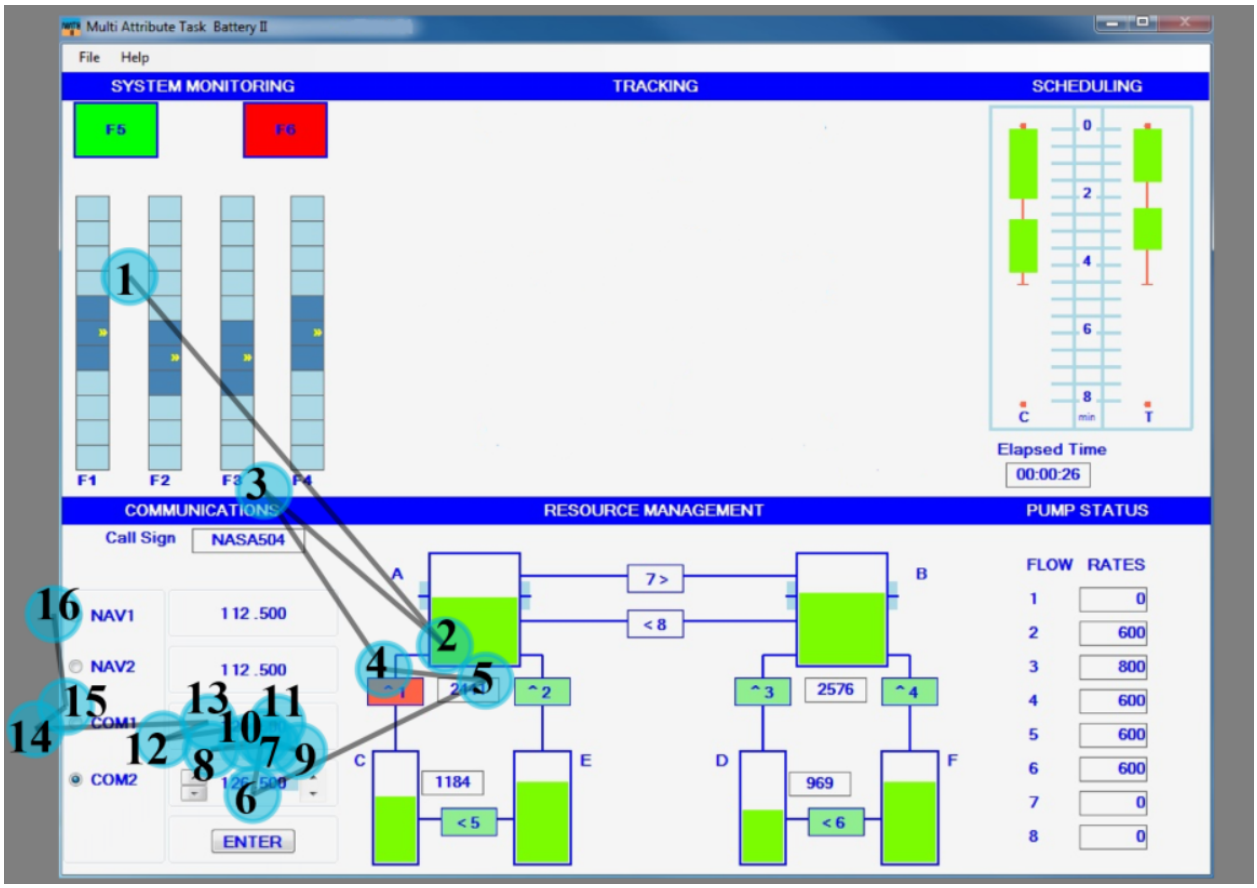


Figure 4.1: Demonstration of fixations and saccades which are important features in characterizing the eye movement data

4.2.2 EEG data

We collected participants' EEG data using a EMOTIV EPOC Flex (128 samples per second sampling rate). In the beginning of the experiments, we recorded participants' baseline EEG measurements with eyes closed and eyes open for one minute each. The device collected EEG signals from 16 channels at the brain according to the 10-20 system, belonging to F,

C and P groups [135]. We then applied a Butterworth band-pass filter at 0.5 Hz and 59 Hz to preprocess the raw EEG data [136, 137, 138]. By applying the Butterworth band-pass filter, further filtration was carried out to obtain Theta (4 to 8 Hz), Alpha (8 to 12 Hz), Beta-low (12 to 18 Hz), Beta-high (18 to 25 Hz), and Gamma (25 to 59 Hz) band data. This preprocessing is done using the **eegkit** package in R.

4.3 Methodology

In this section, we present the feature extraction procedure for both eye movement and EEG data, the rule learning framework, and the decomposition analysis we use to evaluate the multimodal interaction effects discovered from the data.

4.3.1 Feature extraction

Eye movement data

The symptoms associated with ADHD such as impulsiveness, inattention, impaired working memory and executive function deficit have an effect on a subject's eye movement patterns [125]. This relationship between ADHD and eye movement patterns has been studied in previous works [47, 48, 49]. For example, individuals with ADHD may find it difficult to fixate on a particular task or gaze at a particular location for a long time. This can be indicated as differences in "fixation number" and "fixation duration" of individuals with ADHD as compared to those without ADHD. In Figure 4.2, we demonstrate these differences through a visualization of the eye-movement patterns obtained from tracking two participants: one with ADHD and one without ADHD, while they were undertaking the MATB-II tasks. Figures 4.2(a) and 4.2(b) show the visualization of fixation plots obtained by the non-ADHD participant and ADHD participant, respectively. In these figures, the circles show the fixation location, the radius of the circles demonstrates the fixation duration, i.e, fixations of greater duration have larger radii, and the lines joining them show the transitions between fixations. We can see from comparing the two fixation plots that the ADHD participant has a

much larger number of fixation points (fixation number) and the points are also more widely distributed. Moreover, the fixations are of a smaller duration, and a significant number of them lie outside the testing interface, as compared to the non-ADHD participant. While the non-ADHD participant could focus on the software user interface throughout the experiment, the participant with ADHD could not completely focus on the tasks. In addition, Figure 4.2(c) and 4.2(d) show the heat map of fixations for the same two non-ADHD and ADHD participants. We observe the same patterns in the heat map as well. The ADHD participant was less focused and fixated on wider areas on the user interface than the non-ADHD participant. These differences demonstrate the difficulty that individuals with ADHD face due to impaired working memory and executive function, finding it difficult to focus on a particular task or location for a long time. In existing works, a correlation between executive function and some saccadic patterns [139, 140], i.e., rapid ballistic eye movement between two or more fixation points was also found. Further, working memory, which determines the ability to keep information in the mind for brief periods of time, is also related to saccadic patterns and fixation duration [141]. Based on our literature review, we extracted the following eye movement biomarkers: fixation number (characterizes the patterns shown in Figure 4.2(a) and 4.2(b)), saccade velocity, total fixation duration (characterizes the patterns shown in Figure 4.2(c) and 4.2(d)), and left and right pupil diameters. The variables used in the study are then generated by measuring various statistics of these biomarkers including mean, standard deviation, and maximum and minimum values.

EEG data

Electroencephalography (EEG) measures the electrical activity of neurons detected on the scalp. The EEG sensors record brain activity over a specific period of time, where each of the signals in each sensor represents the summed electrical activity occurring in a particular part of the brain. Information extraction from the signal is usually done by measuring the magnitude of oscillations of the signal (power) or change in amplitude of the signal. EEG data has been used to study the group differences in ADHD and non-ADHD subjects [43, 46],

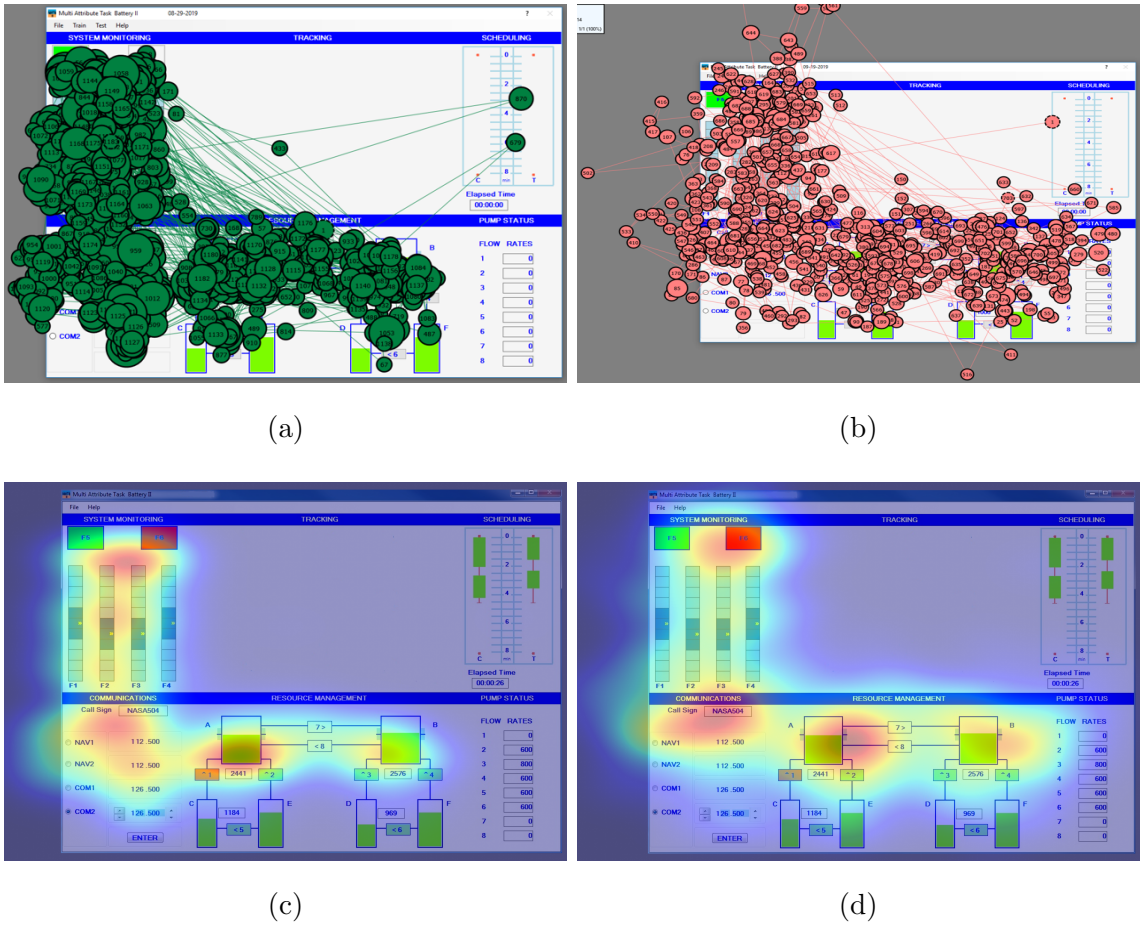


Figure 4.2: Sample plots of (a) fixation plot for non ADHD and (b) fixation plot for ADHD participant, and (c) heat map for non ADHD and (d) heat map for ADHD participant

although its use for ADHD diagnosis has not been reliably established. Nevertheless, several associations between EEG variables and ADHD occurrence have been noted. For example, in one study, it is found that integrating EEG theta/beta ratio (TBR) could help improve diagnostic accuracy from 61% to 88% [44]. In another study that considered TBR and the power of theta and beta bands [142], a 26% increase in TBR in ADHD was observed. Further, [143] consider the fractal dimension, band power, and wavelet coefficients over 22 EEG electrodes and noted that fractal features were more effective than other features

for ADHD classification. [144] presented a semi-supervised feature selection algorithm and classification algorithm using relative theta-beta power, TBR, and theta-alpha ratios of 9 EEG signals to diagnose ADHD and found that power ratio features are dominant features of EEG signal for ADHD. Further, the biomarker theta/beta ratio (TBR) has been found to be relatively lower in the non-ADHD cohort than in ADHD cohort [121]. In several studies, ADHD was associated with increased slow wave activity (theta, 4–8 Hz) or reduced alpha (8–13 Hz) and/or beta activity (13–30 Hz) in the resting EEG [122] as well as during attention task processing [145]. Based on our study of the literature, the following features were extracted from the EEG data: Haguchi fractal dimension (HFD), power spectral density (power), root-mean-square-value (RMSE), and the statistical mean, standard deviation and amplitudes of the signals. Therefore, considering there are 16 channels, each filtered into 5 bands, and we extract 6 features from each band, we have a total of $16 \times 5 \times 6 = 480$ features. For example, F7-alpha-HFD is a feature that refers to the Haguchi fractal dimension of the alpha band of the F7 EEG channel, and similarly, FC5-betaHigh-RMSE refers to the RMSE value of the betaHigh band of the FC5 channel.

4.3.2 The rule learning system

The flowchart of the rule learning system. Fig. 4.3(a) shows the schematic of the rule learning pipeline that we use in our study. Since the number of EEG variables is quite large and more importantly, because these EEG variables are correlated, we used a sparse feature selection model with elastic net regularization [146] to reduce the number of EEG variables by selecting those that are the most representative of the whole set. Elastic net regularization has been shown to be well suited to feature selection in high dimensional EEG data [147, 148, 149]. Using this elastic-net regularization approach, we selected 60 of the 480 EEG variables that were the most significant EEG variables for ADHD prediction. We noticed that these selected features were mostly HFD, RMSE or power features of a few of the channel bands. Then, we combined these 60 EEG features with the 6 features extracted from the eye movement data to obtain a multimodal dataset that we can use to explore

multimodal interaction by the rule learning system.

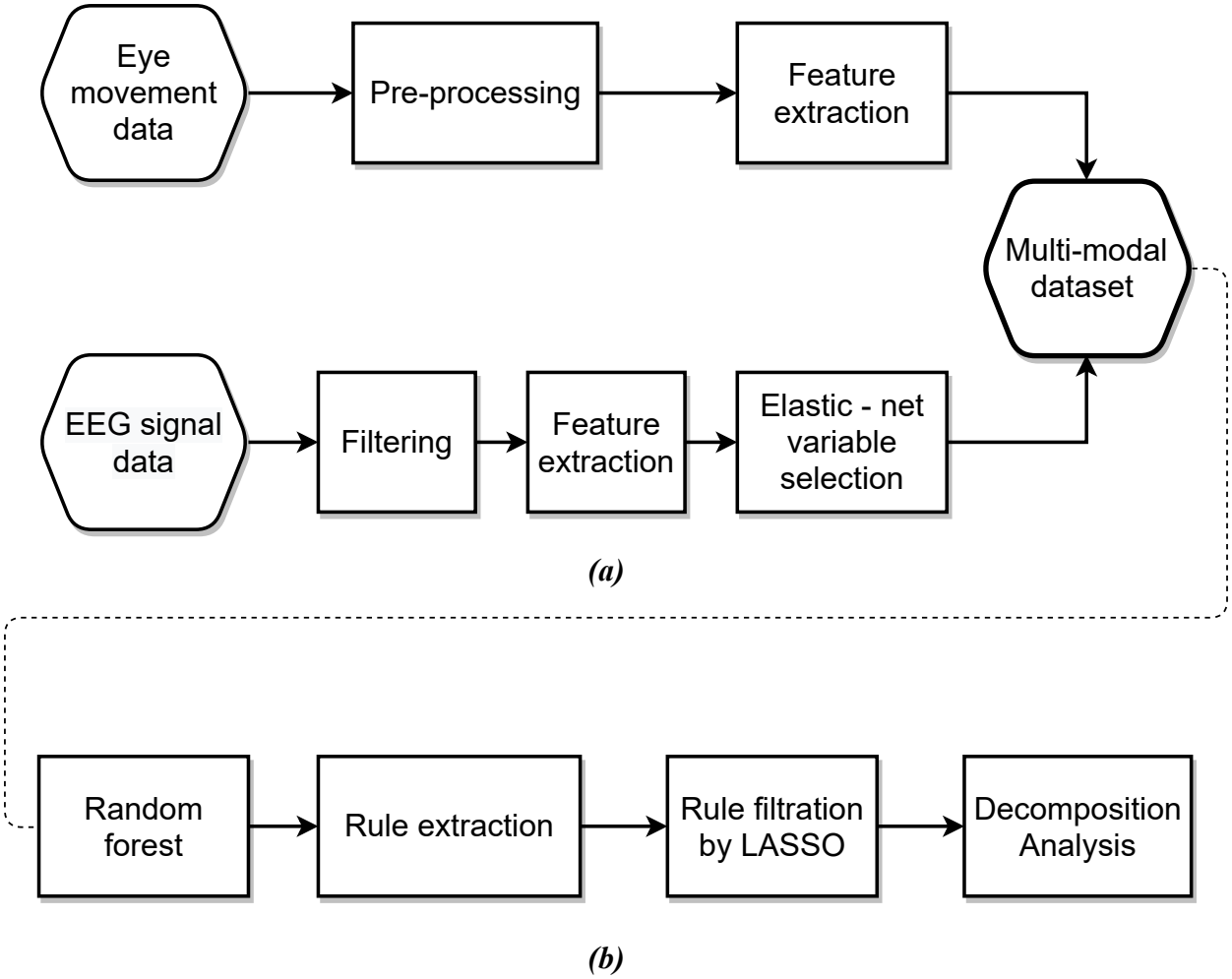
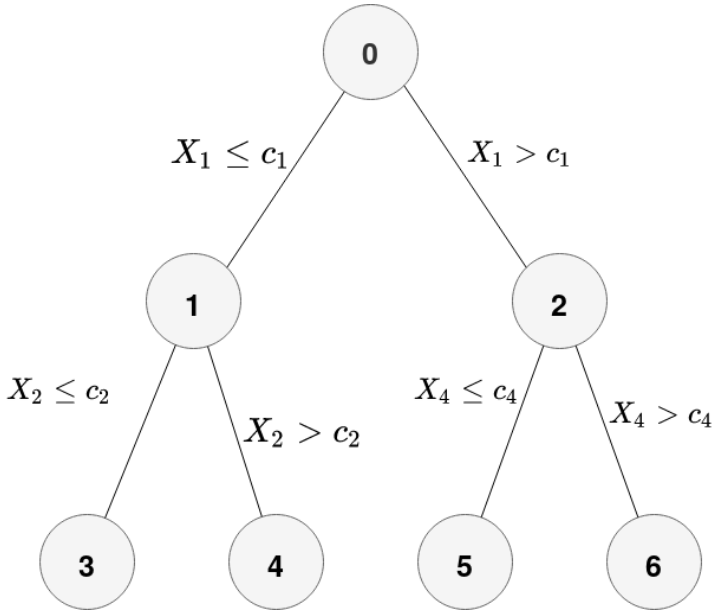


Figure 4.3: ***(a)*** Schematic of data generation model ***(b)*** Schematic of rule extraction and analysis pipeline

The rule learning process. Rules provide an ideal approach to capture interaction effects between variables, as they involve a conjunction of two or more simple conditions about the data. A simple condition restricting the values that an input variable in the data can take can be written $x_j \in s_j$. In other words, this condition restricts the variable x_j to values

belonging to a small subset $s_j \in S_j$ of the domain of x_j. For continuous variables, the condition can also be represented as $p_j \leq x_j \leq q_j$ where p_j and q_j are the lower limit and upper limit of the set s_j, respectively. A rule involves the conjunction of one or more such conditions over different variables to produce a complex, higher dimensional condition of the form $\prod_{i=1}^{N} I(x_j \in s_j)$ where $I()$ is the indicator function that returns 1 if its condition is true, else it returns 0. This rule now contains conditions over multiple variables that must hold simultaneously, thus capturing the interaction effects between them. For example, consider a two-dimensional rule of the form '$x_1 \leq \delta_1$ & $x_2 \geq \delta_2$' that consists of two component variables x_1 and x_2. Say, our outcome variable is affected when this condition holds, which means there is a significant difference in the outcome when the rule holds (conditions on x_1 and x_2 are both true) as compared to when the rule does not hold (either one of the condition is false or both are false). Then it is obvious that this rule is able to capture the interaction between variables x_1 and x_2, and also the location of this interaction that is denoted by the cutoff values δ_1 and δ_2. Note that a rule may consist of more than 2 variables, thus capturing complex multidimensional interaction effects if they exist. Also, the variables in the rule could be bounded on both sides, i.e., by specifying ranges of the variables. Thus, rules provide flexible and natural semantics to represent complex interactions among variables.

The possible number of rules in any given dataset is huge and grows super-exponentially regarding the number of variables and the potential cutoff values. The challenge is how to learn significant rules from data, i.e., rules that have a significant effect on the outcome. Rule learning has always been a main endeavor in AI, and only recently with the development of random forest and sparse learning techniques, efficient rule learning algorithms have been developed [131, 150, 151]. The basic idea of these works, as shown in Fig. 4.3(b), is in two steps: 1) First generate a large set of possible rules using ensemble tree models such as the random forest, and 2) filter this set with sparse regularization methods such as LASSO [152] or elastic net regularization [146]. Following this approach, in our data analysis, we first build a random forest model using both the EEG and eye movement features and gather a set of potential rules. Many of these rules will consist of both EEG variables and eye movement



NODE	EXTRACTED RULES
3	$X_1 \leq c_1$ & $X_2 \leq c_2$
4	$X_1 \leq c_1$ & $X_2 > c_2$
5	$X_1 > c_1$ & $X_4 > c_4$
6	$X_1 > c_1$ & $X_4 > c_4$

Figure 4.4: Decision tree and rules extracted from terminal nodes

variables, capturing the multimodal interactions we hope to identify. We then use LASSO [152] to select the most important rules that can predict the ADHD outcome. The first step, to generate a large set of potential rules from the ensemble tree models has been well documented in our previous works and in the literature so we refer readers to [131, 150, 151] for more details. The idea behind this process is that the decision tree model is a composition of multiple rules, so extracting the rules from a tree is simply to dismember the tree from the root node to leaf nodes. Consider the decision tree and associated table of rules in Fig. 4.4. It can be seen that the path from the root node to each terminal node can be represented as a rule. Therefore, each terminal node of a decision tree can generate a potential rule. To get a large set of rules, we extract such rules from each tree in a random forest trained on our data set. Random forest is an ensemble algorithm that learns many different decision trees by "bagging". Therefore, we can learn a large set of unique rules by extracting rules from each of the trees in a random forest.

Particularly, denote the set of rules extracted from the random forest as $\{r_k\}_1^K$, where K

is the total number of rules, our goal is to select the rules that are most strongly associated with ADHD. To do this, we model a logistic regression between the rules and the ADHD outcome. Assuming N is the number of individuals, denote $\mathbf{X} \in R^{N \times K}$, as the rule matrix where all the elements are binary, i.e., $\mathbf{X}_{ij} = 1$ means the rule r_j applies on the data of individual i, otherwise, $\mathbf{X}_{ij} = 0$. Similarly, $\mathbf{X}_i = \{\mathbf{X}_{i1} \cdots \mathbf{X}_{iK}\}$ is the row vector representing all rule outcomes of individual i. It is easy to see that with a given set of rules extracted from a dataset, it is readily available to translate the original dataset into such a binary rule matrix. After this translation is made, we can now take the rules as the new variables. Let $\mathbf{Y}$ denote the binary outcomes such that $\mathbf{Y}_i = 1$ if individual i belongs to the set of individuals diagnosed with ADHD and $\mathbf{Y}_i = 0$ otherwise. We can then use the outcome variable to guide the selection of rules in $\mathbf{X}$. Specifically, we use the logistic regression model here

$$P(\mathbf{Y}_i = 1 \mid \mathbf{X}_i) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{X}_i)}, \tag{4.1}$$

where $\boldsymbol{\beta}$ are the coefficients of the rules. Using this equation and optimizing for the maximum likelihood provides us the coefficients $\boldsymbol{\beta}$ associated with each of the rules. One way to select the most important rules is to choose rules with the largest coefficients. However, a better way is by using a sparse regularization technique like LASSO [152] which is a technique commonly used for variable selection in high dimensional datasets. It works by penalizing the optimization function by an additional regularization term based on ℓ_1 norm (4.2). LASSO ensures a sparse solution by driving the coefficients of less important rules to zero, thus facilitating in selecting the most important rules. This optimization problem can be written as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; -l(\boldsymbol{\beta}, \mathbf{X}) + \lambda \cdot \|\boldsymbol{\beta}\|_1, \tag{4.2}$$

where $l(\cdot)$ is the log likelihood function of the logistic regression model. We use *Python* programming language for implementing the pipeline. Standard functions from the **scikit-learn** module are used to generate the random forest and run the lasso regression. The rules

are extracted from random forest using our custom code. We will release the implementation code on our GitHub repository.

Decomposition analysis of the rules. It is worthy of mentioning that, being an exploratory method that employs an opportunistic and greedy search over the rule space via random forests and sparse learning, the resulting rules that our algorithm identifies are not necessarily the most *optimal* rules in the data. But on the other hand, if the dataset contains significant multimodal interactions, they are likely to show up in the extracted rules, because of our rule learning system's efficiency and pragmatic efficacy. In other words, our rule learning system is an efficient and efficacious heuristic approach for learning rules from data. Moreover, the learned rules can encode complex meanings, demanding further in-depth analysis of each rule. As our experimental results will show, our method can reveal many of the significant multimodal interactions in the data. Once the rules are extracted, we carry out a decomposition analysis to validate the multimodal interactions. The basic idea of decomposition analysis is simple: since each rule consists of multiple variables, we can prune the rules to remove the redundant variables that do not have a significant effect on the outcome and thus zone in on the true variables contributing to the observed interaction effect. For instance, still consider the rule '$x_1 \leq \delta_1$ & $x_2 > \delta_2$' that consists of two component variables x_1 and x_2. If we knock out x_1 and only keep $x_2 > \delta_2$, we can evaluate the predictive performance of this reduced rule. The same procedure can be done to evaluate the effect of $x_1 \leq \delta_1$. If knocking out a variable causes a huge impact on the predictive performance of the rule, that variable is very important. Or it could happen that either variable alone contributes little to the predictive performance, but when put together, the effect is significantly increased. Decomposition analysis can reveal the roles of the variables in the rule and help us better understand the nature of the interaction. More details of the decomposition analysis and its results will be shown in **Section 4.3**.

4.4 Results

4.4.1 Predictive performance analysis

We built three predictive models following the approach outlined in section 4.3.2: a model only using eye movement data (Model-EM), a model only using EEG data (Model-EEG), and a model using both eye movement and EEG data (Model-both). For each model, we first used the rule learning method to discover the rules, then build a logistic regression model with the rules as variables for ADHD prediction. In Table 4.2, we show the mean model performances of each of the three models by 5-fold cross-validation, along with their standard deviations. It is worth mentioning that because the data is imbalanced, i.e., there are just 12 ADHD observations (13%) among a total of 88 observations, accuracy is not sufficient for model performance comparison, and therefore, we present in total 5 performance measures, namely, the accuracy, balanced accuracy, precision, recall, and F1-score. Accuracy is the proportion of observations correctly identified by the model. Balanced accuracy is the mean of sensitivity and specificity. Sensitivity is the proportion of ADHD observations that were correctly identified, and specificity is the proportion of non-ADHD observations that were correctly identified. Precision is the proportion of predicted ADHD observations that are truly ADHD, while recall (sensitivity) is the proportion of true ADHD observations that were correctly predicted. F1-score is the harmonic mean of precision and recall. It can be seen from Table 4.2 that the 'Model-both' outperforms the other two models in all performance metrics.

4.4.2 Extracted Rules

Table 4.3 shows the top 10 rules extracted from the multimodal dataset using both eye movement and EEG data by our rule learning system described in Section 4.3. Note that the cutoff values of the rules are in their original scale, e.g., fixation duration $>$ 272215.0 in rule 2 means that the sum total of all fixations is greater than 272215.0 milliseconds. In Fig. 4.5, we present the sensitivity of each of the rules, which varies among the rules but

Perf. Measure	**Model-EM**	**Model-EEG**	**Model-both**
Accuracy	0.916 (0.006)	0.943 (0.012)	0.972 (0.006)
Balanced Accuracy	0.827 (0.023)	0.911 (0.027)	0.939 (0.003)
Precision	0.68 (0.016)	0.79 (0.080)	0.892 (0.016)
F1 score	0.68 (0.096)	0.878 (0.053)	0.891 (0.011)
Recall	0.693 (0.049)	0.8 (0.0)	0.88 (0.04)

Table 4.2: Performance comparison of different cases: mean and std. deviation over 5 fold cross-validation.

in general shows the rules are all predictive and can detect ADHD. The specificity of these rules, or the fraction of non-ADHD observations classified as non-ADHD, is found to be 1. This is a common situation in unbalanced datasets where there are very few positives and many negatives.

To gain more understanding of the rules, Fig. 4.6 and Fig. 4.7 aim to visualize their statistical meaning. Geometrically, a rule cuts out a boxed region in the data space, i.e., the regions that contain a red star correspond to the rules. It can be seen how the regions defined by the rules show a pattern towards either ADHD or non-ADHD observations. For example, Fig. 4.6 (a) shows the interaction between F7-alpha-HFD and FC2-theta (std) (i.e., as captured in rule 1), where the regime defined by the interaction contains nearly all the ADHD observations. Similarly, Fig. 4.6 (b) shows interactions between F7-alpha-HFD and left pupil (avg), and in Fig. 4.6 (c), the region defined by the interaction between FC2-betaHigh (std) and FC1-alpha (std) (rule 4) is seen to be containing only ADHD observations, while in Fig. 4.6 (d), the region defined by rule 7 contains only ADHD observations. Similar patterns are observed in Fig. 4.7 as well, i.e., in Fig. 4.7 (a), a potential interaction between F8-theta-HFD and Cz-betaLow-RMSE (i.e., rule 8) is observed, while in (b), the region defined by rule 3 mainly contains ADHD observations, in (c), where the region defined by

the rule 10 is mostly populated by the non-ADHD observations only, and similarly, in (d) a region defined by an interaction between saccade velocity and FC6-gamma-HFD (i.e., rule 9) is also observed to be statistically significant.

Index	**Rule**
1	F7-alpha-HFD> 1.31 & FC2-theta (std)> 1.559 & left pupil (avg)< 2.994
2	F8-gamma-HFD> 2.125 & fixation duration> 272215.0
3	FC2-betaHigh-RMSE> 0.828 & Cz-theta-power< 2625.297 & FC5-betaHigh (std)< 2.757
4	FC2-betaHigh (std)> 0.828 & FC1-alpha (std)< 0.847
5	fixation number> 977.5 & FC5-betaHigh-RMSE> 1.336 & Fz-betaLow-HFD> 1.723
6	fixation duration< 272215.0
7	Pz-alpha (std)< 2.233 & fixation duration< 420219.5
8	F8-theta-HFD> 1.11 & Cz-betaLow-RMSE< 1.421
9	saccade velocity (max)< 0.406 & FC6-gamma-HFD> 2.113 & F8-theta-HFD< 1.116
10	fixation duration> 279332.5 & Fp1-betaLow-HFD> 1.747

Table 4.3: The top 10 rules extracted by our rule learning system

4.4.3 Decomposition analysis

We carried out a decomposition analysis of the rules to understand the roles of the variables in each rule, i.e., some variables may be the most important, and have the greatest contribution to the performance of the rule, while some variables may augment the main variable; in some cases, no single variable alone is sufficient but only when being put together is the synergistic effect strong. To understand these roles, the basic idea of decomposition analysis

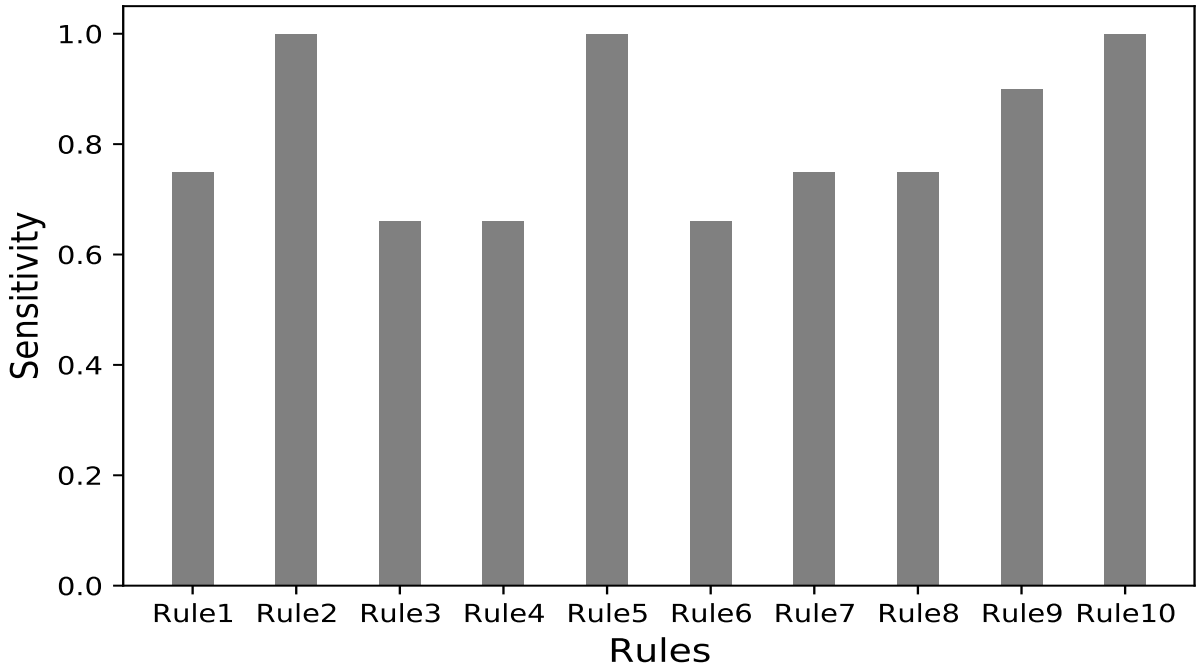


Figure 4.5: Sensitivity of the 10 rules

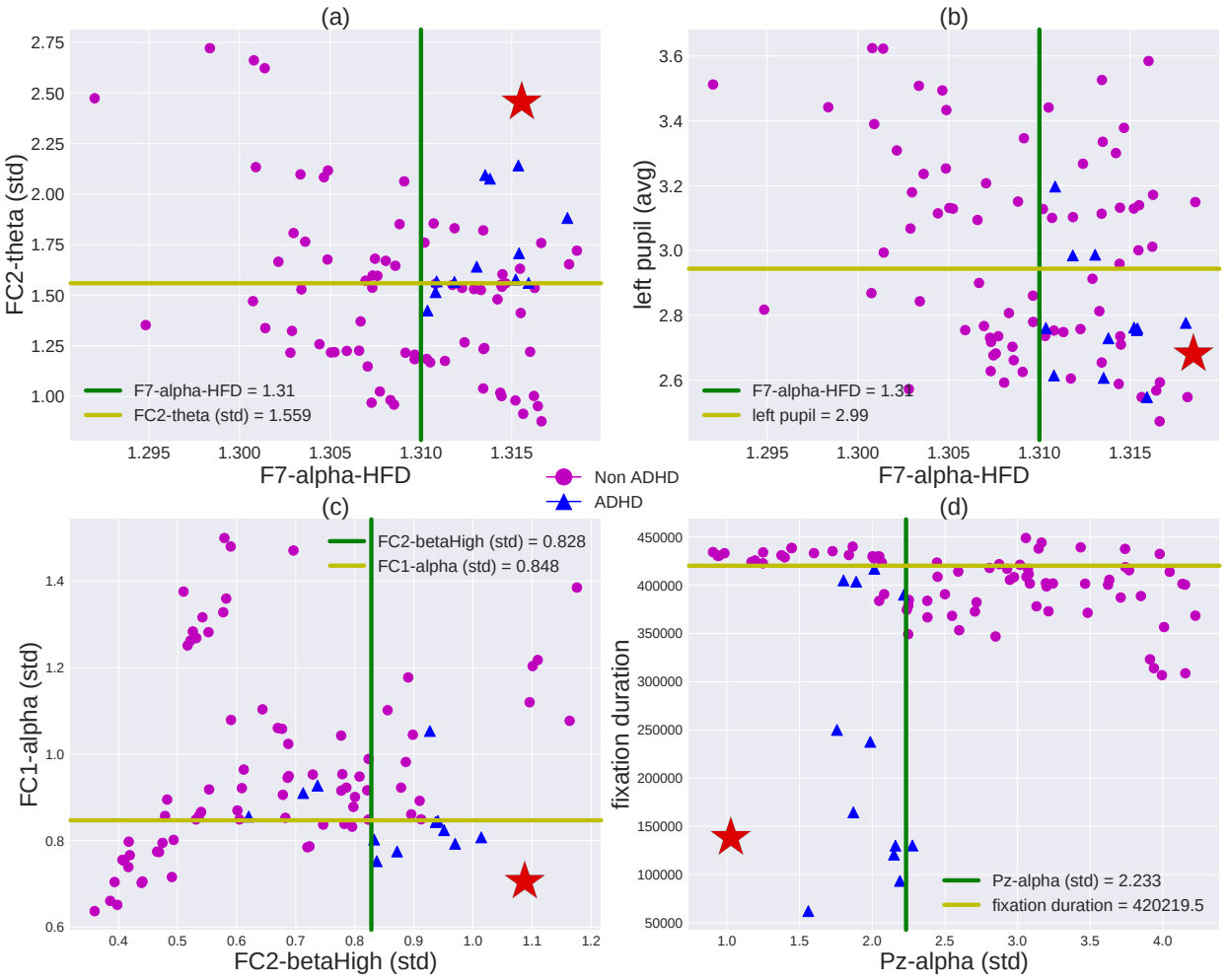


Figure 4.6: Visual exploration of identified interaction effects represented by the rules. Each of the interactions captures the region containing the red star (a) F7-alpha-HFD> 1.31 & FC2-theta (std)> 1.559 (b) F7-alpha-HFD> 1.31 & left pupil (avg)< 2.994 (c) FC2-betaHigh (std)> 0.828 & FC1-alpha (std)< 0.847 (d) Pz-alpha (std)< 2.233 & fixation duration< 420219.5

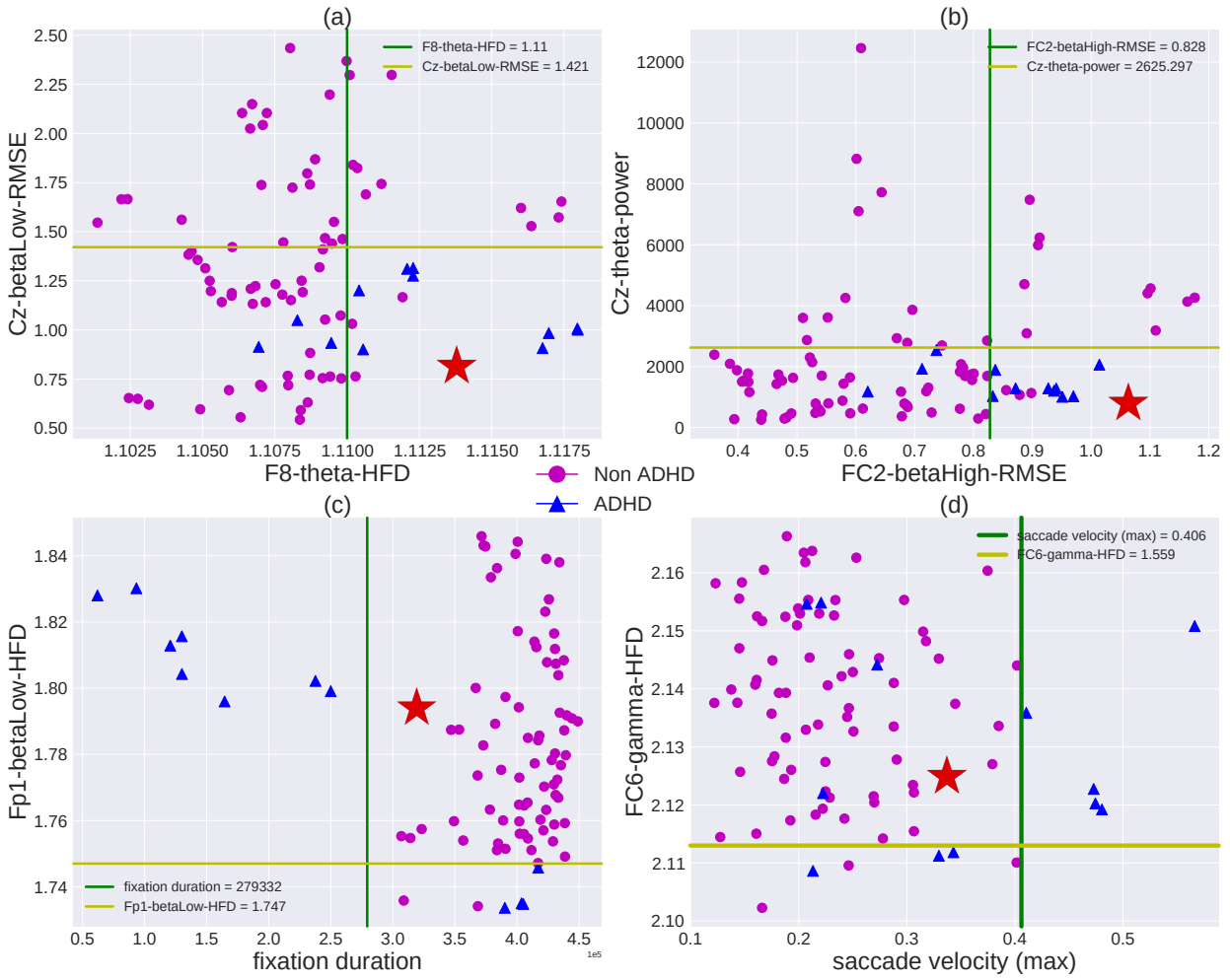


Figure 4.7: Visual exploration of identified interaction effects represented by the rules. Each of the interactions captures the region containing the red star (a) F8-theta-HFD> 1.11 & Cz-betaLow-RMSE< 1.421 (b) FC2-betaHigh-RMSE> 0.828 & Cz-theta-power< 2625.297 (c) fixation duration> 279332.5 & Fp1-betaLow-HFD> 1.747 (d) saccade velocity (max) $\leq$ 0.406 & FC6-gamma-HFD $>$ 2.113

is by "knockout". We begin by first considering the entire rule, i.e., all the variables together, and then proceed by removing one variable from the rule to study the effect of this variable on the performance. This is carried out for each variable separately. Then, we remove each two-variable combination from the rule, and so on depending on the number of variables in the rule until finally, we consider each of the variables taken individually. We evaluate the results of the rule decomposition based on four performance metrics: accuracy, precision, recall and F-score. In Table 4.4 we show the decomposition analysis of some rules presented in Table 4.3, i.e., those which show the interactions between the eye movement and EEG variables. In rule 1, it is clear that the two-variable interactions outperform the single factors, and the full rule that consists of all the 3 variables is the best according to the accuracy. This decomposition analysis indicates that there are strong statistical effects among all the 3 variables. Particularly, it can be seen that although the two variables, FC2-theta (std) and left pupil (avg), are only marginally predictive if considered alone, when it is combined together the predictive performance is much improved. This is a solid indication that there is an interaction among the two variables. The interpretation of rule 2, however, tells a different story. In rule 2, the interaction between F8-gamma-HFD & fixation duration shows a marginal decrease in accuracy, i.e., from 0.95 to 0.93, but there is a significant increase in recall from 0.67 to 1.0. In unbalanced datasets such as our case here where ADHD observations only form a small proportion of the dataset, predicting the small number of true positives is much more important. Thus, a high recall metric is very valuable. A similar trend holds in rule 5. In the interaction between fixation number & FC5-betaHigh-rmse where though accuracy decreases from 0.89 to 0.85, an increase in recall is observed from 0.42 to 0.75 by considering the interaction. Other interactions that are clearly exposed by decomposition analysis are in rule 7 between Pz-alpha (std) & fixation duration, where all the metrics are greatly improved by considering both variables in tandem. Similarly, the interaction between saccade velocity & F8-theta-HFD in rule 9 shows a much larger recall value than each of the variables taken individually. These results suggest that analyzing the EEG and eye movement data in a multimodal model not only helps us improve the predictive

performance of the models but also discovers some deeper physiological causes and effects of the condition.

Table 4.4: Decomposition analysis of multimodal interactions

N	No. feat.	Decomposition Analysis	Acc.	Prec.	Recall	F-score
1	all	F7-alpha-HFD> 1.31 & FC2-theta (std)> 1.559 & left pupil (avg)< 2.994	0.92	0.69	0.75	0.72
	2	F7-alpha-HFD> 1.31 & left pupil (avg)≤ 2.994	0.8	0.39	0.92	0.55
		FC2-theta (std)> 1.559 & left pupil (avg)≤ 2.994	0.84	0.45	0.75	0.56
	1	F7-alpha-HFD> 1.31	0.4	0.0	0.0	0.0
		FC2-theta (std)> 1.559	0.65	0.26	0.83	0.39
		left pupil (avg)≤ 2.994	0.44	0.03	0.08	0.04

Continuation of Table 4.4

N	No. feat.	Decomposition Analysis	Acc.	Prec.	Recall	F-score
2	all	F8-gamma-HFD> 2.12 & fix. duration> 272215	0.93	0.67	1.0	0.8
	1	F8-gamma-HFD> 2.12	0.84	0.4	0.33	0.36
		fix. duration> 272215	0.95	1.0	0.67	0.8
5	all	fix. number > 977.5 & FC5-betaHigh-rmse> 1.336 & Fz-betaLow-HFD > 1.723	0.89	0.55	1	0.71
	2	fix. number > 977.5 & FC5-betaHigh-rmse> 1.336	0.85	0.47	0.75	0.58
		fix. number > 977.5 & Fz-betaLow-HFD > 1.723	0.92	0.73	0.67	0.7
	1	fix. number > 977.5	0.89	0.62	0.42	0.5
		FC5-betaHigh-rmse > 1.336	0.82	0.33	0.33	0.33
		Fz-betaLow-HFD > 1.723	0.91	1.0	0.33	0.5
7	all	Pz-alpha(std)≤2.23 & fix. duration≤ 420219.5	0.97	0.85	0.92	0.88
	1	Pz-alpha(std)≤2.23	0.73	0.32	0.92	0.48
		fix. duration≤ 420219.5	0.52	0.0	0.0	0.0
9	all	saccade vel.(max)≤0.406 & FC6-gamma-HFD> 2.113 & F8-theta-HFD≤1.116	0.91	0.61	0.92	0.73
	2	saccade vel.(max)≤0.406 & FC6-gamma-HFD> 2.113	0.92	0.73	0.67	0.7
		saccade vel.(max)≤ 0.406 & F8-theta-HFD≤1.116	0.91	0.67	0.67	0.67

Continuation of Table 4.4

N	No. feat.	Decomposition Analysis	Acc.	Prec.	Recall	F-score
	1	saccade vel.(max)≤0.406	0.92	1.0	0.42	0.59
		FC6-gamma-HFD> 2.113	0.86	0.5	0.25	0.33
		F8-theta-HFD≤ 1.116	0.86	0.5	0.33	0.4

Continuation of Table 4.4

N	No. feat.	Decomposition Analysis	Acc.	Prec.	Recall	F-score
10	all	fix. duration> 279332.5 & Fp1-betaLow-HFD>	0.98	0.86	1.0	0.92
	1	fix. duration> 279332.5	0.95	1.0	0.67	0.8
		Fp1-betaLow-HFD> 1.747	0.89	0.67	0.33	0.44

4.4.4 Validation

We further use the permutation test [153] to validate the statistical reliability and significance of our data analysis results. The basic idea is to randomly assign labels to the training data, repeat the whole data analysis process described in our paper, and see what prediction performance is achieved. While assigning the random labels, we made sure to maintain the relative frequency of the labels in the dataset. If the same prediction performance can be achieved on this randomized dataset, then it indicates that our data analysis process has the potential to overfit the data and therefore the results may not be reliable. But on the other hand, if the same prediction performance is not achieved, then it suggests our data analysis results are solid. Specifically, in our analysis, we first create a randomized dataset by altering the training dataset and randomly assigning labels to observations while keeping the relative frequency of the labels constant. Then we run our method on the altered dataset and obtain the cross-validated classification performance. We repeat this process 100 times and the results are shown in Table 4.5. We also plot a histogram of the cross-validated accuracy's obtained from the randomized labels for each of these 100 trials, as well as the original cross-validated accuracy in Figure 4.8.

It can be observed that the classification performance on this randomized dataset is much lower than the classification performance we have obtained on the original dataset on all five

Perf. Measure	Original Labels	Flipped Labels
Accuracy	0.972 (0.006)	0.735 (0.047)
Balanced Accuracy	0.939 (0.003)	0.514 (0.082)
Precision	0.892 (0.016)	0.127 (0.099)
F1 score	0.891 (0.011)	0.129 (0.097)
Recall	0.88 (0.04)	0.176 (0.130)

Table 4.5: Performance comparison of model using original labels and randomized labels (Mean and std. deviation)

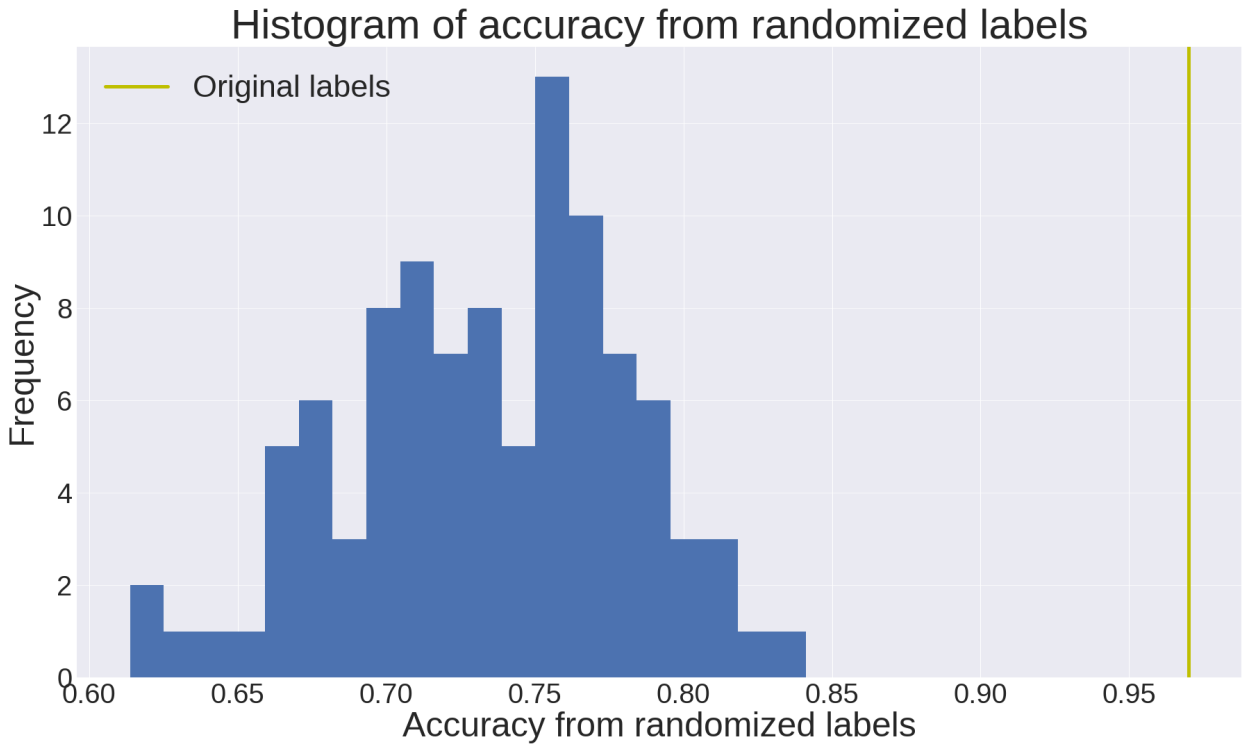


Figure 4.8: Histogram of accuracy obtained by training on randomized labels and original labels

measures. While accuracy is relatively high due to the imbalance in the dataset, the precision and recall is much lower. Thus, the permutation method shows that our original model was able to capture statistically significant effects.

4.5 Discussion

In this work, we conducted an exploratory analysis using a rule learning and analysis framework to identify multimodal biomarkers of eye movement and EEG data for ADHD prediction. Our rule learning approach is similar to RuleFit [131] but our contribution lies on our integrated design of the learning and analysis framework that consists of data integration, reduction, rule learning, and decomposition analysis. It is important to understand that rule learning is an exploratory approach that yields a list of high-quality rules, but these rules are not exclusive or globally optimal. Thus, we conduct decomposition analysis to evaluate the statistical strength of the interactions of the biomarkers in the rules. Our study showed not only better predictive performance using multimodal data, but also revealed interesting interactions among eye movement biomarkers with EEG biomarkers that may suggest new hypotheses for further studies that connect different physiological aspects of the same disease.

Multimodal integration of EEG and eye tracking has recently been studied for emotion recognition [154, 155], aesthetic appreciation [156], cognitive workload [157]. However, there has been a lack of studies for ADHD. To the best of our knowledge, our work is the first that aims to identify multimodal biomarkers for ADHD prediction. The performance comparison of the results shown in Table 4.2 demonstrates that the multimodal biomarkers could lead to better prediction, i.e., on each of the 5 different performance measures considered, using both eye movement and EEG data leads to significantly better performance as compared to the single-modality models. We obtain an accuracy of 0.916 from a model built using eye movement data, 0.943 using EEG data, while a combined dataset gives a significantly higher accuracy of 0.972. Similarly, the model built on eye movement data provides a recall of 0.69, 0.8 for EEG data, but a recall of 0.88 is achieved by the model using both modalities

of biomarkers. The precision of the eye movement model is 0.68, EEG model is 0.79, but using both datasets gives a precision of 0.89. Likewise, F1 score of the eye movement model is 0.68, EEG model is 0.87, but using both is 0.89. These results demonstrate clearly that a multimodal model provides better results.

Decomposition analysis done in Table 4.4 quantifies the interactions between the eye movement and EEG variables by comparing the classification performance of the decomposed interactions using 4 measures: accuracy, precision, recall and F1 score. The findings in Table 4.4 also show higher prediction accuracy when both modalities are combined than single modality models. We also identified interesting interactions between these two modalities that may further lead to follow-up studies of the mechanisms of these interactions. For example, in rule 1, we observe that the interaction between F7-alpha-HFD and left pupil (avg) is significant, as when the variables are combined in a rule (denoted with *), the accuracy, precision, recall, and F-score measures are higher than each of them taken individually. Likewise, there is also an interaction between FC2-theta (std) and left pupil (avg). Similarly, in rule 2, F8-gamma-HFD and fixation duration combined together could lead to a recall of 1, showing that the interaction is able to detect a much greater number of ADHD observations than either of the variables taken individually (0.33 and 0.67), while the accuracy at 0.93 is also very high. This is important because in unbalanced datasets, the sensitivity is a critical measure, apart from accuracy. In rule 3, a significant interaction is detected between Pz-alpha (std) and fixation duration, and the combined rule shows much higher performance on all measures compared to the individual case. In rule 8, interactions are detected between saccade velocity (max) and FC6-gamma-HFD and saccade velocity (max) and F8-theta-HFD.

Several of the significant variables that we observed in our rules have been detected in prior works, which further demonstrate the efficacy of our rules in detecting clinically meaningful patterns. For example, we see in rules 6 and 10 that a shorter fixation duration is associated with ADHD, i.e., as shown in the fixation gaze plot in Fig. 4.2a and 4.2c. Significantly shorter fixation duration among the ADHD group has indeed been noted in prior works [124, 158]. It is hypothesized that fixation duration is related to processing

rate [159] and increases when processing becomes more effort-demanding, particularly when greater attention is deployed to the fixation location [160]. Similar to our rules 2, 6 and 7, ADHD group has also been observed to have reduced complexity in full band EEG HFD calculations [161] while in particular beta and then gamma band HFD were found to be most discriminative of major depressive disorder. Beta band EEG activity has been shown to be related to attention modulation, and decreases in beta activity in elderly subjects reflect deficits in sustaining attention processing [162]. [163], a commonly co-morbidity condition with ADHD. In Fig. 4.2a and 4.2c we visualize the gaze plot of a non-ADHD participant and an ADHD participant and observe that the ADHD participant has a smaller fixation duration as well as a larger number of fixations. This trend is also observed in rule 5 where ADHD in individuals is associated with higher fixation numbers. Prior studies have identified increased micro-saccade rates in individuals with ADHD traits, [164, 139] which suggests higher fixation numbers for ADHD individuals. Rule 5 also shows an association of higher RMSE of the beta-High band with ADHD, suggesting an interaction between the higher fixation number. Increased beta activity in boys with ADHD has been noted in literature [165], particularly in temporal lobes, but also in the Fz channel. This could explain the interaction obtained in rule 5.

Other discovered interactions characterized by the rules are also worthy of further investigation. For example, in rule 1, we see that higher pupil size is associated with ADHD, which has been also reported in previous works. A linear relationship was observed between pupil diameter and several measures of task performance that suggested that attentional lapses occurred when pupil diameter was small [166]. Frontal-midline (like FC2) theta oscillations have been suggested as the "working language of executive functioning" linked to cognitive processing and performance, which could show up as the interaction effect. Furthermore, it was reported that pupil size is [167] can be related to hyperactive/impulsive symptoms associated with hyper-responsiveness [167, 168] with higher pupil dilation linked to happy faces. Since theta band activity has been linked to emotional processing in several studies [169], it indicates that there might be a physiological mechanism mediating the interaction

effect between the pupil size and theta band activity as characterized by rule 1.

Our study is an exploratory study that has its own limitations. First, the dataset we collected is of a small sample size. Particularly, the number of ADHD subjects is much smaller than the control subjects, with only 12 out of 88 observations being ADHD. The robustness of the rule learning method on a small sample size would alleviate this problem, and the further decomposition analysis provides a conservative evaluation of the discovered rules. But still, it is important to note that our results are suggestive of new hypotheses but not confirmatory. Our results also show that there is significant value to collect both eye movement data and EEG data simultaneously for ADHD studies. In the future, we will expand on the data collection and validate the rules we found in this study. Particularly, we will recruit more ADHD subjects and test the performances of these rules on the new dataset. A second limitation is that the ADHD diagnosis was self-reported by our participants, instead of being directly confirmed by a physician or clinic. On the other hand, the use of self-reported ADHD has been standard practice in existing ADHD studies. For instance, the prevalence of self-reported ADHD has been studied in [170] where participants were asked whether they were clinically diagnosed with ADHD. Likewise, self-reported scores on a scale of ADHD symptoms have also been used [171]. In other studies, the parent-reported ADHD variable has been used to leverage machine learning techniques to identify predictors of receiving psychosocial treatment for ADHD [172] and identified parent-reported ADHD severity (mild vs. moderate/severe) as the factor that best predicts which children receive psychosocial treatment. Further, in studies [173, 174] that used machine learning algorithms to distinguish behaviors of ADHD and autism spectrum disorder (ASD), self-reported diagnosis of ADHD was also used. Further, studies [170, 171] show that this self-reported ADHD was associated with significant psychiatric comorbidity and is a clinically valid and effective variable of ADHD outcome. Thus, our study is built on a solid outcome variable. However, we do plan to collect more data in the future with a more comprehensive plan including diagnostic interviews conducted by psychiatrists to build a more powerful prediction model that could be potentially used in clinical practices. Another limitation is the use of a relatively simple

prediction model in this study to combine the rules, i.e., after the rules are identified, a logistic regression model is used with the rules as the variables to predict ADHD. When we have more data in the future, more sophisticated prediction models will be built to further enhance the model prediction accuracy. Recently, a rule-based prediction model has been developed that can model heterogeneity in complex biomedical datasets for accurate and sequential predictions [151] along with uncertainty quantification capacity. It is well known that ADHD is a highly heterogeneous disorder characterized by differing clinical profiles, patterns of cognitive disabilities, developmental trajectories, and a wide range of functional brain anomalies [175]. Considering this heterogeneity and complexity of ADHD, and the need to prevent misdiagnoses, our multimodal rule-based approach may be enhanced to model the uncertainty in ADHD prediction as well as the heterogeneous effects. Moreover, in this study, limited by the sample size, we only focused on a few EEG and eye movement biomarkers. In the future, we will include more biomarkers in the rule-learning process that may help us identify more rules.

4.6 Conclusion

In this study, we conducted an exploratory data analysis that leverages on recent developments in rule learning to identify interesting interactions among two modalities of data, the eye movement data and EEG signal data, for ADHD prediction. Compared with existing works on predictive models of ADHD that have mostly focused on a single modality of data, our approach is more integrative and achieves better predictive performance than models built on a single modality alone. Further, the multimodal biomarkers we identified yield not only better performance, but also discover interactions among the two modalities which would lead to the generation of interesting new hypotheses about the pathology of ADHD.

Chapter 5

CONCLUSION AND FUTURE RESEARCH

In this dissertation, we presented novel rule-based statistical and machine learning methods that seek to address some of the challenges and opportunities put forward by modern medical datasets such as interpretability, variable sparsity, and heterogeneity. Further, we also presented an application of rule learning in a study for the discovery of multimodal biomarkers for ADHD.

5.1 Rule learning, analysis and heterogeneity

We presented a rule learning and analysis framework to tackle the challenge of multidimensionality in rule learning for survival analysis. Our method, SURVFIT, is the first of its kind that we know of to address the challenge of both rule sparsity and variable sparsity in rule extraction. To further the analyses of the rules, we presented a rule analysis framework that involves statistical testing, decomposition analysis and cutoff sensitivity. We also developed an R package, SURVFIT, implementing our rule extraction and analysis framework. Next, we presented a Bayesian semi-parametric framework to generate ordered rule lists for heterogeneous survival data. Our framework uses ordered rule lists to model the heterogeneity in datasets by dividing it into different subgroups, while also being capable of quantifying uncertainty in the predictions. We have applied our proposed methods to the study of a real-world sepsis survival dataset with convincing computational results that demonstrate the validity and potential application of our approach.

Future work: Future directions in rule learning research could explore further applications of group sparsity norms to obtain structured solutions. For example, applications such as genomics studies with known and pre-existing group structures among the different

genes may be modeled using such an approach. Another area of research can be the development of efficient algorithms, using alternative loss functions for survival analysis such as the more standard Cox proportional hazards regression. The rule learning methods are not only interpretable, but can also be concise in the suggestions they make about the underlying process. Therefore, we suggest extending this work to other applications where trustworthy and communicable decisions are imperative, such as the study of other conditions such as Type 1 Diabetes or Alzheimers.

5.2 Discovery of multimodal biomarkers

We presented an application of rule-learning to discover multimodal biomarkers for the diagnosis of ADHD in a study that sought to combine the information from EEG and eye-tracking data for the first time. Compared to existing approaches, that focused on the use of a single modality of data, our approach was more integrative and achieved better performance than models built on a single modality alone. Finally, we were able to validate many of our results through prior works in the study of both sepsis and ADHD.

Future work: Future work can build on the promise of integrating eye movement and EEG data for ADHD diagnosis. Particularly, larger and more representative datasets may be collected to enhance the reliability of results. Further, collecting the EEG and eye-movement data while subjects are carrying out the particular tests used to evaluate ADHD, rather than a generic multitasking test, may be used for better performance. Adding more modalities of data, such as test scores, or diagnostic reports from interviews conducted by psychiatrists could be used to build a more powerful model that may be used in clinical practice.

BIBLIOGRAPHY

[1] Hans van Houwelingen and Hein Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

[2] Jie Hao, Youngsoon Kim, Tejaswini Mallavarapu, Jung Hun Oh, and Mingon Kang. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC medical genomics*, 12(10):1–13, 2019.

[3] Sijia Huang, Nicole Chong, Nathan E Lewis, Wei Jia, Guoxiang Xie, and Lana X Garmire. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome medicine*, 8(1):1–14, 2016.

[4] Adrian E Raftery, David Madigan, and Chris T Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics*, 5:323–349, 1996.

[5] Giorgio Paulon, Maria De Iorio, Alessandra Guglielmi, and Francesca Ieva. Joint modeling of recurrent events and survival: a Bayesian non-parametric approach. *Biostatistics*, 2020.

[6] Sushil Mittal, David Madigan, Randall S. Burd, and Marc A. Suchard. High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics*, 15(2):207–221, April 2014.

[7] Desmond YH Yap, Colin SO Tang, Maggie KM Ma, Man Fai Lam, and Tak Mao Chan. Survival analysis and causes of mortality in patients with lupus nephritis. *Nephrology Dialysis Transplantation*, 27(8):3248–3254, 2012.

[8] Wanqing Chen and Rongshou Zheng. Incidence, mortality and survival analysis of breast cancer in china. *Chinese Journal of Clinical Oncology*, (13):668–674, 2015.

[9] Kai He, Shuai Huang, and Xiaoning Qian. Early detection and risk assessment for chronic disease with irregular longitudinal data analysis. *Journal of biomedical informatics*, 96:103231, 2019.

[10] Symeon E Christodoulou. Water network assessment and reliability analysis by use of survival analysis. *Water Resources Management*, 25(4):1229–1238, 2011.

[11] Gabriel KR Pereira, Luís F Guilardi, Kiara S Dapieve, Cornelis J Kleverlaan, Marilia P Rippe, and Luiz Felipe Valandro. Mechanical reliability, fatigue strength and survival analysis of new polycrystalline translucent zirconia ceramics for monolithic restorations. *Journal of the mechanical behavior of biomedical materials*, 85:57–65, 2018.

[12] Tina Harrison and Jake Ansell. Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing*, 6(3):229–239, 2002.

[13] Richard E Barlow and Frank Proschan. Statistical theory of reliability and life testing: probability models. Technical report, Florida State Univ Tallahassee, 1975.

[14] Jadzia Cendrowska. PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4):349–370, October 1987.

[15] William W. Cohen. Fast Effective Rule Induction. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 115–123. Morgan Kaufmann, San Francisco (CA), January 1995.

[16] R. S. Michalski. Pattern Recognition as Rule-Guided Inductive Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(4):349–361, July 1980.

[17] J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5(3):239–266, August 1990.

[18] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[19] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[20] Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, September 2008. arXiv: 0811.1679.

[21] Ying Lin, Xiaoning Qian, Jeffrey Krischer, Kendra Vehik, Hye-Seung Lee, and Shuai Huang. A Rule-Based Prognostic Model for Type 1 Diabetes by Identifying and Synthesizing Baseline Profile Patterns. *PLOS ONE*, 9:e91095, 2014.

[22] B. M. Patil, R. C. Joshi, and D. Toshniwal. Association Rule for Classification of Type-2 Diabetic Patients. In *2010 Second International Conference on Machine Learning and Computing*, pages 330–334, February 2010.

[23] Enrico Glaab, Jaume Bacardit, Jonathan M. Garibaldi, and Natalio Krasnogor. Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data. *PLOS ONE*, 7(7):e39932, July 2012.

[24] Ying Wu, Shuai Huang, and Xiangyu Chang. Understanding the complexity of sepsis mortality prediction via rule discovery and analysis: a pilot study. *BMC medical informatics and decision making*, 21(1):1–15, 2021.

[25] Marjolein Fokkema. Fitting Prediction Rule Ensembles with R Package pre. *arXiv:1707.07149 [stat]*, July 2017. arXiv: 1707.07149.

[26] Łukasz Wróbel, Adam Gudyś, and Marek Sikora. Learning rule sets from survival data. *BMC Bioinformatics*, 18, May 2017.

[27] Mark Robert Segal. Regression Trees for Censored Data. *Biometrics*, 44(1):35–47, 1988.

[28] M. LeBlanc and J. Crowley. Relative risk trees for censored survival data. *Biometrics*, 48(2):411–425, June 1992.

[29] Odd O Aalen. Heterogeneity in survival analysis. *Statistics in medicine*, 7(11):1121–1137, 1988.

[30] Luc Duchateau and Paul Janssen. *The frailty model*. Springer Science & Business Media, 2007.

[31] Cécile Proust-Lima, Mbéry Séne, Jeremy MG Taylor, and Hélène Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research*, 23(1):74–90, 2014.

[32] Mona Haghighi, Suzanne Bennett Johnson, Xiaoning Qian, Kristian F. Lynch, Kendra Vehik, Shuai Huang, and The TEDDY Study Group. A Comparison of Rule-based Analysis with Regression Methods in Understanding the Risk Factors for Study Withdrawal in a Pediatric Study. *Scientific Reports*, 6:30828, 2016.

[33] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[34] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106:1039–1082, 2017.

[35] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian Treed Models. *Machine Learning*, 48:299–320, 2002.

[36] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9:1350–1371, 2015.

[37] Saeed Amal, Lida Safarnejad, Jesutofunmi A Omiye, Ilies Ghanzouri, John Hanson Cabot, and Elsie Gyang Ross. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine*, 9, 2022.

[38] Andrzej W Przybyszewski, Mark Kon, Stanislaw Szlufik, Artur Szymanski, Piotr Habela, and Dariusz M Koziorowski. Multimodal learning and intelligent prediction of symptom development in individual parkinson's patients. *Sensors*, 16(9):1498, 2016.

[39] Wei Shao, Tongxin Wang, Liang Sun, Tianhan Dong, Zhi Han, Zhi Huang, Jie Zhang, Daoqiang Zhang, and Kun Huang. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. *Medical image analysis*, 65:101795, 2020.

[40] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

[41] Kevin M Boehm, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P Shah. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2):114–126, 2022.

[42] Yuhui Du, Xingyu He, Peter Kochunov, Godfrey Pearlson, L Elliot Hong, Theo GM van Erp, Aysenil Belger, and Vince D Calhoun. A new multimodality fusion classification approach to explore the uniqueness of schizophrenia and autism spectrum disorder. *Human brain mapping*, 43(12):3887–3903, 2022.

[43] Agatha Lenartowicz and Sandra K Loo. Use of eeg to diagnose adhd. *Current psychiatry reports*, 16(11):498, 2014.

[44] Steven M Snyder, Thomas A Rugino, Mady Hornig, and Mark A Stein. Integration of an eeg biomarker with a clinician's adhd evaluation. *Brain and behavior*, 5(4):e00330, 2015.

[45] Mohammad Reza Mohammadi, Ali Khaleghi, Ali Moti Nasrabadi, Safa Rafieivand, Moslem Begol, and Hadi Zarafshan. Eeg classification of adhd and normal children using non-linear features and neural network. *Biomedical Engineering Letters*, 6(2):66–73, 2016.

[46] Sandra K. Loo and Russell A. Barkley. Clinical utility of eeg in attention deficit hyperactivity disorder. *Applied Neuropsychology*, 12(2):64–76, 2005. PMID: 16083395.

[47] Astar Lev, Yoram Braw, Tomer Elbaum, Michael Wagner, and Yuri Rassovsky. Eye tracking during a continuous performance test: Utility for assessing adhd patients. *Journal of Attention Disorders*, page 1087054720972786, 2020.

[48] Belgüzar Nilay Türkan, Sonia Amado, Eyüp Sabri Ercan, and Ipek Perçinel. Comparison of change detection performance and visual search patterns among children with/without adhd: Evidence from eye movements. *Research in developmental disabilities*, 49:205–215, 2016.

[49] Randal G Ross, Josette G Harris, Ann Olincy, and Allen Radant. Eye movement task measures inhibition and spatial working memory in adults with schizophrenia, adhd, and a normal comparison group. *Psychiatry Research*, 95(1):35–42, 2000.

[50] Bhaskar Sen, Neil C Borle, Russell Greiner, and Matthew RG Brown. A general prediction model for the detection of adhd and autism using structural and functional mri. *PloS one*, 13(4):e0194856, 2018.

[51] Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao, and Bao-Liang Lu. Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3071–3074, 2019.

[52] Russell A Barkley and Mary Jo Poillion. Attention deficit hyperactivity disorder: a handbook for diagnosis and treatment. *Behavioral disorders*, 19(2):150–152, 1994.

[53] Ying Lin, Shuai Huang, Gregory E. Simon, and Shan Liu. Data-based Decision Rules to Personalize Depression Follow-up. *Scientific Reports*, 8, 2018.

[54] Mona Haghighi, Amanda Smith, Dave Morgan, Brent Small, and Shuai for the Alzheimer's Disease Neuroimaging Initiative (ADNI) Huang. Identifying Cost-Effective Predictive Rules of Amyloid-β Level by Integrating Neuropsychological Tests and Plasma-Based Markers. *Journal of Alzheimer's Disease*, 43(4):1261–1270, December 2014.

[55] Torsten Hothorn and Berthold Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137, 2003.

[56] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008.

[57] Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, March 2017.

[58] Robert Tibshirani. The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16(4):385–395, 1997.

[59] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, February 2006.

[60] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, April 2013.

[61] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[62] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, April 2013.

[63] Chuyang Ke, Yan Jin, Heather Evans, Bill Lober, Xiaoning Qian, Ji Liu, and Shuai Huang. Prognostics of surgical site infections using dynamic health data. *Journal of Biomedical Informatics*, 65:22–33, January 2017.

[64] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured Variable Selection with Sparsity-Inducing Norms. *J. Mach. Learn. Res.*, 12:2777–2824, November 2011.

[65] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, June 2012.

[66] Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, January 2009.

[67] Lei Yuan, Jun Liu, and Jieping Ye. Efficient Methods for Overlapping Group Lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2104–2116, September 2013.

[68] Jun Liu, Shuiwang Ji, and Jieping Ye. *SLEP: Sparse Learning with Efficient Projections*. 2009.

[69] Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966. Publisher: Pacific Journal of Mathematics.

[70] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[71] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[72] Aliasghar Tarkhan and Noah Simon. Bigsurvsgd: Big survival data analysis via stochastic gradient descent, 2020.

[73] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50:163–170, 1966.

[74] J Martin Bland and Douglas G Altman. The odds ratio. *BMJ*, 320(7247):1468, 2000.

[75] Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247:2543–2546, 1982.

[76] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

[77] Jeffrey E Gotts and Michael A Matthay. Sepsis: pathophysiology and clinical management. *BMJ*, 353, 2016.

[78] Vincent Liu, Gabriel J. Escobar, John D. Greene, Jay Soule, Alan Whippy, Derek C. Angus, and Theodore J. Iwashyna. Hospital Deaths in Patients With Sepsis From 2 Independent CohortsHospital Deaths in Patients With SepsisLetters. *JAMA*, 312(1):90–92, July 2014.

[79] Rui P. Moreno, Barbara Metnitz, Leopold Adler, Anette Hoechtl, Peter Bauer, Philipp G. H. Metnitz, and SAPS 3 Investigators. Sepsis mortality prediction based on predisposition, infection and response. *Intensive Care Medicine*, 34(3):496–504, March 2008.

[80] Dee W Ford, Andrew J Goodwin, Annie N Simpson, Emily Johnson, Nandita Nadig, and Kit N Simpson. A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data. *Critical care medicine*, 44(2):319–327, February 2016.

[81] David Andaluz-Ojeda, Felipe Bobillo, Verónica Iglesias, Raquel Almansa, Lucía Rico, Francisco Gandía, Salvador Resino, Eduardo Tamayo, Raul Ortiz de Lejarazu, and Jesús F. Bermejo-Martin. A combined score of pro- and anti-inflammatory interleukins improves mortality prediction in severe sepsis. *Cytokine*, 57(3):332–336, March 2012.

[82] Timothy E. Sweeney, Thanneer M. Perumal, Ricardo Henao, Marshall Nichols, Judith A. Howrylak, Augustine M. Choi, Jesús F. Bermejo-Martin, Raquel Almansa, Eduardo Tamayo, Emma E. Davenport, Katie L. Burnham, Charles J. Hinds, Julian C. Knight, Christopher W. Woods, Stephen F. Kingsmore, Geoffrey S. Ginsburg, Hector R. Wong, Grant P. Parnell, Benjamin Tang, Lyle L. Moldawer, Frederick E. Moore, Larsson Omberg, Purvesh Khatri, Ephraim L. Tsalik, Lara M. Mangravite, and Raymond J. Langley. A community approach to mortality prediction in sepsis via gene expression analysis. *Nature Communications*, 9(1):694, February 2018.

[83] Supreeth P Shashikumar, Matthew D Stanley, Ismail Sadiq, Qiao Li, Andre Holder, Gari D Clifford, and Shamim Nemati. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *Journal of electrocardiology*, 50(6):739–743, 2017.

[84] Marc Leone, Sami Blidi, François Antonini, Bertrand Meyssignac, Sébastien Bordon, Frédéric Garcin, Aude Charvet, Valéry Blasco, Jacques Albanèse, and Claude Martin. Oxygen Tissue Saturation Is Lower in Nonsurvivors than in Survivors after Early Resuscitation of Septic Shock. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 111(2):366–371, August 2009. Publisher: The American Society of Anesthesiologists.

[85] Nicolas Nesseler, Yoann Launey, Caroline Aninat, Fabrice Morel, Yannick Mallédant,

and Philippe Seguin. Clinical review: The liver in sepsis. *Critical Care*, 16(5):235, 2012.

[86] Jiaying Dou, Yiping Zhou, Yun Cui, Min Chen, Chunxia Wang, and Yucai Zhang. AST-to-Platelet Ratio Index as Potential Early-Warning Biomarker for Sepsis-Associated Liver Injury in Children: A Database Study. *Frontiers in Pediatrics*, 7, 2019. Publisher: Frontiers.

[87] Jessica A. Zagory, Avafia Dossa, Jamie Golden, Aaron R. Jensen, Catherine J. Goodhue, Jeffrey S. Upperman, and Christopher P. Gayer. Re-evaluation of liver transaminase cutoff for CT after pediatric blunt abdominal trauma. *Pediatric Surgery International*, 33(3):311–316, March 2017.

[88] P. Krafft, H. Steltzer, M. Hiesmayr, W. Klimscha, and A. F. Hammerle. Mixed venous oxygen saturation in critically ill septic shock patients. The role of defined events. *Chest*, 103(3):900–906, March 1993.

[89] Jennifer V. Pope, Alan E. Jones, David F. Gaieski, Ryan C. Arnold, Stephen Trzeciak, and Nathan I. Shapiro. Multicenter Study of Central Venous Oxygen Saturation (ScvO2) as a Predictor of Mortality in Patients With Sepsis. *Annals of Emergency Medicine*, 55(1):40–46.e1, January 2010.

[90] Yong Yang, Kok Soong Yang, Yin Maw Hsann, Vincent Lim, and Biauw Chi Ong. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *Journal of Critical Care*, 25(3):398–405, September 2010.

[91] Tran Dd, Groeneveld Ab, van der Meulen J, Nauta Jj, Strack van Schijndel Rj, and Thijs Lg. Age, chronic disease, sepsis, organ system failure, and mortality in a medical intensive care unit. *Critical Care Medicine*, 18(5):474–479, May 1990.

[92] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical*

learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.

[93] Houtao Deng. Interpreting Tree Ensembles with inTrees. *arXiv:1408.5456 [cs, stat]*, August 2014. arXiv: 1408.5456.

[94] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[95] Imad Bou-Hamad, Denis Larocque, Hatem Ben-Ameur, et al. A review of survival trees. *Statistics surveys*, 5:44–71, 2011.

[96] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

[97] Sara Martino, Rupali Akerkar, and Håvard Rue. Approximate Bayesian Inference for Survival Models. *Scandinavian Journal of Statistics*, 38:514–528, 2011.

[98] Heikki Joensuu, Aki Vehtari, Jaakko Riihimäki, Toshirou Nishida, Sonja E. Steigen, Peter Brabec, Lukas Plank, Bengt Nilsson, Claudia Cirilli, Chiara Braconi, Andrea Bordoni, Magnus K. Magnusson, Zdenek Linke, Jozef Sufliarsky, Massimo Federico, Jon G. Jonasson, Angelo Paolo Dei Tos, and Piotr Rutkowski. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The Lancet. Oncology*, 13:265–274, 2012.

[99] Ying Lin, Shuai Huang, Gregory E Simon, and Shan Liu. Analysis of depression trajectory patterns using collaborative learning. *Mathematical biosciences*, 282:191–203, 2016.

[100] Ying Lin, Shan Liu, and Shuai Huang. Selective sensing of a heterogeneous population of units with dynamic health conditions. *IISE Transactions*, 50(12):1076–1088, 2018.

[101] Samy S Abu-Naser and Bashar G Bastami. A proposed rule based system for breasts cancer diagnosis. *WWJMRD*, 2016.

[102] Anthony N. Nguyen, Michael J. Lawley, David P. Hansen, Rayleen V. Bowman, Belinda E. Clarke, Edwina E. Duhig, and Shoni Colquist. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*, 17:440–445, 2010.

[103] Evangelos J. Giamarellos-Bourboulis, Anna Norrby-Teglund, Vassiliki Mylona, Athina Savva, Iraklis Tsangaris, Ioanna Dimopoulou, Maria Mouktaroudi, Maria Raftogiannis, Marianna Georgitsi, Anna Linnér, George Adamis, Anastasia Antonopoulou, Efterpi Apostolidou, Michael Chrisofos, Chrisostomos Katsenos, Ioannis Koutelidakis, Katerina Kotzampassi, George Koratzanis, Marina Koupetori, Ioannis Kritselis, Korina Lymberopoulou, Konstantinos Mandragos, Androniki Marioli, Jonas Sundén-Cullberg, Anna Mega, Athanassios Prekates, Christina Routsi, Charalambos Gogos, Carl-Johan Treutiger, Apostolos Armaganidis, and George Dimopoulos. Risk assessment in sepsis: a new prognostication rule by APACHE II score and serum soluble urokinase plasminogen activator receptor. *Critical Care*, 16:R149, 2012.

[104] James E. Barrett and Anthony C. C. Coolen. Gaussian process regression for survival data with competing risks. *arXiv:1312.1591*, 2014.

[105] Robert B Gramacy and Herbert K. H Lee. Bayesian Treed Gaussian Process Models With an Application to Computer Modeling. *Journal of the American Statistical Association*, 103:1119–1130, 2008.

[106] Hyoung-Moon Kim, Bani K Mallick, and CC Holmes. Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668, 2005.

[107] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for ma-*

chine learning. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006.

[108] Virgilio Gómez-Rubio and Håvard Rue. Markov chain Monte carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051, 2018.

[109] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[110] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. *JMLR*, 2015.

[111] Alan D Saul, James Hensman, Aki Vehtari, and Neil D Lawrence. Chained Gaussian processes. In *Artificial Intelligence and Statistics*, pages 1431–1440, 2016.

[112] GPy. GPy: A Gaussian process framework in Python. `http://github.com/SheffieldML/GPy`, since 2012.

[113] M Ibrahim, PWC Prasad, Abeer Alsadoon, and L Pham. Synchronous virtual classroom for student with adhd disorder. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6. IEEE, 2016.

[114] Christopher A Meyers and Richard G Bagnall. A case study of an adult learner with asd and adhd in an undergraduate online learning environment. *Australasian Journal of Educational Technology*, 31(2), 2015.

[115] George J DuPaul, Thomas J Power, Arthur D Anastopoulos, and Robert Reid. *ADHD Rating Scale—IV: Checklists, norms, and clinical interpretation.* Guilford Press, 1998.

[116] João Ricardo Sato, Marcelo Queiroz Hoexter, André Fujita, and Luis Augusto Rohde. Evaluation of pattern recognition and feature extraction methods in adhd prediction. *Frontiers in systems neuroscience*, 6:68, 2012.

[117] Jason W Bohland, Sara Saperstein, Francisco Pereira, Jérémy Rapin, and Leo Grady. Network, anatomical, and non-imaging measures for the prediction of adhd diagnosis in individual subjects. *Frontiers in systems neuroscience*, 6:78, 2012.

[118] Shashank Jaiswal, Michel F. Valstar, Alinda Gillott, and David Daley. Automatic detection of adhd and asd from expressive behaviour in rgbd data. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 762–769, 2017.

[119] Ortal Slobodin, Inbal Yahav, and Itai Berger. A machine-based prediction model of adhd using cpt data. *Frontiers in human neuroscience*, 14:383, 2020.

[120] Akira Yasumura, Mikimasa Omori, Ayako Fukuda, Junichi Takahashi, Yukiko Yasumura, Eiji Nakagawa, Toshihide Koike, Yushiro Yamashita, Tasuku Miyajima, Tatsuya Koeda, et al. Applied machine learning method to predict children with adhd using prefrontal cortex activity: a multicenter study in japan. *Journal of attention disorders*, 24(14):2012–2020, 2020.

[121] Martijn Arns, C Keith Conners, and Helena C Kraemer. A decade of eeg theta/beta ratio research in adhd: a meta-analysis. *Journal of attention disorders*, 17(5):374–383, 2013.

[122] Robert J Barry, Adam R Clarke, Stuart J Johnstone, Christopher A Magee, and Jacqueline A Rushby. Eeg differences between eyes-closed and eyes-open resting conditions. *Clinical neurophysiology*, 118(12):2765–2773, 2007.

[123] Gavindya Jayawardena, Anne Michalek, and Sampath Jayarathna. Eye tracking area of interest in the context of working memory capacity tasks. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 208–215. IEEE, 2019.

[124] Pamela Deans, Liz O'Laughlin, Brad Brubaker, Nathan Gay, Damon Krug, et al. Use of eye movement tracking in the differential diagnosis of attention deficit hyperactivity disorder (adhd) and reading disability. *Psychology*, 1(04):238, 2010.

[125] Senuri De Silva, Sanuwani Dayarathna, Gangani Ariyarathne, Dulani Meedeniya, Sampath Jayarathna, Anne MP Michalek, and Gavindya Jayawardena. A rule-based system for adhd identification using eye movement data. In *2019 Moratuwa Engineering Research Conference (MERCon)*, pages 538–543. IEEE, 2019.

[126] Yan Jin, Yi Su, Xiao-Hua Zhou, Shuai Huang, and Alzheimer's Disease Neuroimaging Initiative. Heterogeneous multimodal biomarkers analysis for alzheimer's disease via bayesian network. *EURASIP Journal on Bioinformatics and Systems Biology*, 2016:1–8, 2016.

[127] Raja Parasuraman. Neuroergonomics: Brain, cognition, and performance at work. *Current directions in psychological science*, 20(3):181–186, 2011.

[128] Ranjana K Mehta and Raja Parasuraman. Neuroergonomics: a review of applications to physical and cognitive work. *Frontiers in human neuroscience*, 7:889, 2013.

[129] JC Woestenburg, MN Verbaten, and JL Slangen. The removal of the eye-movement artifact from the eeg by regression analysis in the frequency domain. *Biological psychology*, 16(1-2):127–147, 1983.

[130] Michael Plöchl, José Pablo Ossandón, and Peter König. Combining eeg and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in human neuroscience*, 6:278, 2012.

[131] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[132] Ameer Hamza Shakur, Shuai Huang, Xiaoning Qian, and Xiangyu Chang. Survfit: Doubly sparse rule learning for survival data. *Journal of Biomedical Informatics*, 117:103691, 2021.

[133] Ameer Hamza Shakur, Tianchen Sun, Ji-Eun Kim, and Shuai Huang. A rule-based exploratory analysis for discovery of multimodal biomarkers of adhd using eye movement and eeg data. *IISE Transactions on Healthcare Systems Engineering*, pages 1–15, 2022.

[134] Yamira Santiago-Espada, Robert R Myer, Kara A Latorella, and James R Comstock Jr. The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide. 2011.

[135] Manzoor Khazi, Atul Kumar, and MJ Vidya. Analysis of eeg using 10: 20 electrode system. *International Journal of Innovative Research in Science, Engineering and Technology*, 1(2):185–191, 2012.

[136] Atsuo Murata and Hirokazu Iwase. Analysis of chaotic dynamics in eeg and its application to assessment of mental workload. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)*, volume 3, pages 1579–1582. IEEE, 1998.

[137] Seung-Hyeon Oh, Yu-Ri Lee, and Hyoung-Nam Kim. A novel eeg feature extraction method using hjorth parameter. *International Journal of Electronics and Electrical Engineering*, 2(2):106–110, 2014.

[138] Sushil Chandra, Greeshma Sharma, Mansi Sharma, Devendra Jha, and Alok Pakash Mittal. Workload regulation by sudarshan kriya: an eeg and ecg perspective. *Brain informatics*, 4(1):13–25, 2017.

[139] Maria Panagiotidi, Overton Paul, and Stafford Tom. Increased microsaccade rate in individuals with adhd traits. *Journal of Eye Movement Research*, 10(1), 2017.

[140] Yusuke Goto, Kazuo Hatakeyama, Toshihiro Kitama, Yu Sato, Hideaki Kanemura, Kakuro Aoyagi, Kanji Sugita, and Masao Aihara. Saccade eye movements as a quantitative measure of frontostriatal network in children with adhd. *Brain and Development*, 32(5):347–355, 2010.

[141] Martijn J Schut, Nathan Van der Stoep, Albert Postma, and Stefan Van der Stigchel. The cost of making an eye movement: A direct link between visual working memory and saccade execution. *Journal of Vision*, 17(6):15–15, 2017.

[142] Geir Ogrim, Juri Kropotov, and Knut Hestad. The quantitative eeg theta/beta ratio in attention deficit/hyperactivity disorder and normal controls: sensitivity, specificity, and behavioral correlates. *Psychiatry research*, 198(3):482–488, 2012.

[143] Khadijeh Sadatnezhad, Reza Boostani, and Ahmad Ghanizadeh. Classification of bmd and adhd patients using their eeg signals. *Expert Systems with Applications*, 38(3):1956–1963, 2011.

[144] Berdakh Abibullaev and Jinung An. Decision support algorithm for diagnosis of adhd using electroencephalograms. *Journal of medical systems*, 36(4):2675–2688, 2012.

[145] Islam F Halawa, Basma B El Sayed, Omnia R Amin, Nagwa A Meguid, and Ann A Abdel Kader. Frontal theta/beta ratio changes during tova in egyptian adhd children. *Neurosciences Journal*, 22(4):287–291, 2017.

[146] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[147] Mayrim Vega-Hernández, Eduardo Martínez-Montes, José M Sánchez-Bornot, Agustín Lage-Castellanos, and Pedro A Valdés-Sosa. Penalized least squares methods for solving the eeg inverse problem. *Statistica Sinica*, pages 1535–1551, 2008.

[148] Li-Chen Shi and Bao-Liang Lu. Eeg-based vigilance estimation using extreme learning machines. *Neurocomputing*, 102:135–143, 2013.

[149] Deirel Paz-Linares, Mayrim Vega-Hernandez, Pedro A Rojas-Lopez, Pedro A Valdes-Hernandez, Eduardo Martinez-Montes, and Pedro A Valdes-Sosa. Spatio temporal eeg source imaging with the hierarchical bayesian elastic net and elitist lasso models. *Frontiers in neuroscience*, 11:635, 2017.

[150] Marjolein Fokkema. Fitting prediction rule ensembles with r package pre. *arXiv preprint arXiv:1707.07149*, 2017.

[151] Ameer Hamza Shakur, Xiaoning Qian, Zhangyang Wang, Bobak Mortazavi, and Shuai Huang. Gpsrl: Learning semi-parametric bayesian survival rule lists from heterogeneous patient data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10608–10615. IEEE, 2021.

[152] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[153] Polina Golland, Feng Liang, Sayan Mukherjee, and Dmitry Panchenko. Permutation tests for classification. In *International conference on computational learning theory*, pages 501–515. Springer, 2005.

[154] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and eeg to enhance emotion recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[155] Hao Tang, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep neural networks. In *International Conference on Neural Information Processing*, pages 811–819. Springer, 2017.

[156] Fu Guo, Mingming Li, Mingcai Hu, Fengxiang Li, and Bozhao Lin. Distinguishing and quantifying the visual aesthetics of a product: an integrated approach of eye-tracking and eeg. *International Journal of Industrial Ergonomics*, 71:47–56, 2019.

[157] Jesus L Lobo, Javier Del Ser, Flavia De Simone, Roberta Presta, Simona Collina, and Zdenek Moravek. Cognitive workload classification using eye-tracking and eeg data. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*, pages 1–8, 2016.

[158] Canan Karatekin and Robert F Asarnow. Exploratory eye movements to pictures in childhood-onset schizophrenia and attention-deficit/hyperactivity disorder (adhd). *Journal of Abnormal Child Psychology*, 27(1):35–49, 1999.

[159] Kirk Moffitt. Evaluation of the fixation duration in visual search. *Perception & Psychophysics*, 27(4):370–372, 1980.

[160] Radha Nila Meghanathan, Cees van Leeuwen, and Andrey R Nikolaev. Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Frontiers in human neuroscience*, 8:1063, 2015.

[161] Miray Altınkaynak, Nazan Dolu, Ayşegül Güven, Ferhat Pektaş, Sevgi Özmen, Esra Demirci, and Meltem İzzetoğlu. Diagnosis of attention deficit hyperactivity disorder with combined time and frequency features. *Biocybernetics and Biomedical Engineering*, 40(3):927–937, 2020.

[162] Mateusz Gola, Mikołaj Magnuski, Izabela Szumska, and Andrzej Wróbel. Eeg beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects. *International Journal of Psychophysiology*, 89(3):334–341, 2013.

[163] Mehran Ahmadlou, Hojjat Adeli, and Amir Adeli. Fractality analysis of frontal brain in major depressive disorder. *International Journal of Psychophysiology*, 85(2):206–211, 2012.

[164] Todd D Gould, Theresa M Bastain, Margaret E Israel, Daniel W Hommer, and F Xavier Castellanos. Altered performance on an ocular fixation task in attention-deficit/hyperactivity disorder. *Biological psychiatry*, 50(8):633–635, 2001.

[165] Ching-Tai Chiang, Chen-Sen Ouyang, Rei-Cheng Yang, Rong-Ching Wu, and Lung-Chang Lin. Increased temporal lobe beta activity in boys with attention-deficit hyperactivity disorder by loreta analysis. *Frontiers in Behavioral Neuroscience*, 14:85, 2020.

[166] Ruud L van den Brink, Peter R Murphy, and Sander Nieuwenhuis. Pupil diameter tracks lapses of attention. *PLoS One*, 11(10):e0165274, 2016.

[167] Johan Lundin Kleberg, Matilda A Frick, and Karin C Brocki. Increased pupil dilation to happy faces in children with hyperactive/impulsive symptoms of adhd. *Development and psychopathology*, 33(3):767–777, 2021.

[168] Moshe Fried, Eteri Tsitsiashvili, Yoram S Bonneh, Anna Sterkin, Tamara Wygnanski-Jaffe, Tamir Epstein, and Uri Polat. Adhd subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision research*, 101:62–72, 2014.

[169] Qian Luo, Xi Cheng, Tom Holroyd, Duo Xu, Frederick W Carver, and James Blair. Theta band activity in response to emotional expressions and its relationship with gamma band activity as revealed by meg and advanced beamformer source imaging. *Frontiers in human neuroscience*, 7:940, 2014.

[170] Jacqueline Hesson and Ken Fowler. Prevalence and correlates of self-reported add/adhd

in a large national sample of canadian adults. *Journal of Attention Disorders*, 22(2):191–200, 2018.

[171] Laura M Garnier-Dykstra, Gillian M Pinchevsky, Kimberly M Caldeira, Kathryn B Vincent, and Amelia M Arria. Self-reported adult attention-deficit/hyperactivity disorder symptoms among college students. *Journal of American College Health*, 59(2):133–136, 2010.

[172] Anne S Morrow, Alexandro D Campos Vega, Xin Zhao, and Michelle M Liriano. Leveraging machine learning to identify predictors of receiving psychosocial treatment for attention deficit/hyperactivity disorder. *Administration and Policy in Mental Health and Mental Health Services Research*, 47(5):680–692, 2020.

[173] M Duda, R Ma, N Haber, and DP Wall. Use of machine learning for behavioral distinction of autism and adhd. *Translational psychiatry*, 6(2):e732–e732, 2016.

[174] M Duda, N Haber, J Daniels, and DP Wall. Crowdsourced validation of a machine-learning classification system for autism and adhd. *Translational psychiatry*, 7(5):e1133–e1133, 2017.

[175] Yuyang Luo, Dana Weibman, Jeffrey M Halperin, and Xiaobo Li. A review of heterogeneity in attention deficit/hyperactivity disorder (adhd). *Frontiers in human neuroscience*, 13:42, 2019.