Intelligent Assistance for Expert-Driven Subpopulation Discovery in High-Dimensional Time-Stamped Medical Data

Uli Niemann

18.01.2021

Contents

1 Introduction **4**

1.1 Motivation and Objectives 4

1.2 Structure and Contributions of This Thesis 5

2 Medical Background & Datasets **7**

2.1 Fundamental terms and study types 7

2.2 Basic research 8

2.3 Clinical studies 8

2.4 Epidemiological studies 9

2.5 Challenges 9

2.6 Medical Applications 9

I SUBPOPULATION DISCOVERY IN HIGH-DIMENSIONAL DATA **16**

3 Interactive Discovery and Inspection of Subpopulations **17**

3.1 Motivation and Comparison to Related Work 17

3.2 Subpopulation Discovery Workflow and Interactive Mining Assistant 21

3.3 Experiments and Findings 29

3.4 Conclusions on … 35

4 Identifying Distinct Subpopulations **36**

4.1 Motivation and Comparison to Related Work 37

4.2 Finding Distinct Classification Rules 39

4.3 Experimental Setup 41

4.4 Results 41

4.5 Conclusions on identifying distinct subpopulations 45

5 Visual Identification of Informative Features 46

5.1 Motivation and Comparison to Related Work 47
5.2 Selection of Measurement Instruments 47
5.3 Identification of Tinnitus Phenotypes using Clustering 48
5.4 Visualisation of Phenotypes 49
5.5 Interactive Components and Exploration of Treatment Effects 53
5.6 Results 53
5.7 Clinical Value 57
5.8 Discussion 57
5.9 Conclusions 58

II EXPLOITING DYNAMICS 60

6 Constructing Evolution Features to Capture Study Participant Change over Time 61

6.1 Motivation and Comparison to Related Work 61
6.2 Evolution Features 63
6.3 Evaluation Setup 65
6.4 Results 67
6.5 Conclusions from Exploiting Study Participant Evolution 70

7 Feature Extraction From Short Temporal Sequences for Clustering 71

III POST-MINING FOR INTERPRETATION 72

8 Post-Hoc Interpretation of Classification Models 73

8.1 Motivation and Methodological Underpinnings 74
8.2 Overview of the Mining Workflow 78
8.3 Results 85
8.4 Discussion 95
8.5 Conclusion 100

9 Subpopulation-Specific Learning and Post-Hoc Model Interpretation 101

9.1 Motivation and Comparison to Related Work 101
9.2 Comparing Differences in Feature Importance between Two Subpopulations . . . 101
9.3 Workflow 102
9.4 Results 102
9.5 Conclusions on Subpopulation-Specific Differences in Feature Importance 102

IV SUMMARY **103**

10 Conclusion and Future Work **104**

10.1 Research Results for Medical Expert-Guided Knowledge Discovery 104

10.2 Future Work 104

Abbreviations **105**

A Variables selected for phenotyping **106**

Chapter 1

Introduction

1.1 Motivation and Objectives

Common objectives of data analysis in medical research include (i) identifying long-term determinants and protective factors of a medical condition of interest, (ii) discovering subpopulations with increased outcome prevalence, and (iii) generating robust statistical models that can explain relationships between one or more independent variables and the target variable. For example, epidemiologists attempt to discover associations between multiple risk factors and an outcome in cohort studies by collecting data that include extensive information about participants obtained from questionnaires, medical examinations, laboratory analyses, and imaging. Often these data are collected repeatedly over time, for example in longitudinal studies. This means that there is latent but often overlooked temporal information in such data, the investigation of which can potentially lead to new insights.

To find associations between variables, medical researchers usually first carefully derive some hypotheses from clinical practice, experimental studies, or extensive literature reviews to then test them formally for statistical significance. However, with the ever-increasing volume and heterogeneity of medical data, traditional hypothesis-driven workflows are becoming increasingly impractical, as for this reason some important inherent associations between variables may go undetected. Machine learning can improve medical research by discovering understandable descriptions of patient or study participant subpopulations that are similar in outcome, and thus can be used to derive new hypotheses.

The proliferation of medical machine learning applications is triggered by several reasons, such as the desire to make automated use of the plethora of information collected about study subjects, but sometimes just based on the ubiquity of deep learning success stories in the media. However, the ease of creating complex data-driven models is no guarantee that insights can be effortlessly derived. Most state-of-the-art machine learning algorithms such as deep neural networks and gradient boosting machines generate so-called black-box models with multiple layers of complexity that involve many multivariate, nonlinear interactions between variables that are difficult to represent intuitively. It is critical that the application expert, who is not a practitioner but a scientist working in a clinical or epidemiological setting, be equipped with tools to understand, explore, and visualize the models so that they can drill down to specific individual patterns and gain actionable insights that ultimately contribute to prevention, diagnosis, and treatment in clinical practice. Because medical data come from a wide variety of sources key characteristics of the collected datasets vary, requiring adaptation of methods to the specifics of each application scenario.

The goal of our work is to develop methods that serve as intelligent assistance to medical researchers in the analysis of high-dimensional, temporal medical data. Hence, the core research question of the thesis is: How to derive accurate yet understandable patterns for subpopulation discovery in high-dimensional temporal medical data? Both before, during, and after the generation of machine learning models, several challenges must be overcome in order for the domain expert to be able to derive actionable knowledge. These challenges can be translated into the following four requirements. (1) *Comprehensibility of patterns*: the extracted models, including clusters, rules, and other patterns, must be made understandable; preferably, the model generation process is also comprehensible. (2) *Exploitation of time*: latent temporal information must be exploited, while satisfying requirement 1. (3) *Minimization of redundancy*: redundancy must be minimized, while satisfying requirement (1). Patterns may overlap in terms of the topics they cover. This leads to a redundancy of patterns, which has a negative impact on the perceived quality of the model. Our task is to extract, process and display the most relevant (temporal) patterns for expert-driven model exploration.

1.2 Structure and Contributions of This Thesis

This thesis presents solutions to support medical researchers in the data-driven analysis of high-dimensional, temporal medical data. Design decisions and developments were partly inspired by suggestions from the respective domain experts and cooperation partners, including a specialist for internal medicine, an epidemiologist with statistical expertise, a diabetes expert and three tinnitus specialists. The thesis is organized into three parts and ten chapters tackling the aforementioned research question and challenges. PART I covers methods for discovering subpopulations in high-dimensional data. PART II focuses specifically on temporal aspects of medical datasets and provides approaches that extract informative representations from latent temporal data. PART III addresses post-hoc analysis of machine learning models and includes solutions to derive model-, observation-, and subpopulation-level insights from otherwise "opaque" black boxes.

- Chapter 2 (*Medical Background*) presents relevant medical background information, a brief comparison of medical study types and an overview of the case studies our solutions are tailored to.
- Chapter 3 (*Interactive Discovery and Inspection of Subpopulations*) presents a workflow for the **TODO**
- Chapter 4 (*Identifying Distinct Subpopulations*) examines redundancy in large rule sets describing subpopulations. We present a workflow that extracts a smaller number of representative rules. These rules are selected to avoid instance overlap as much as possible, thus covering different concepts in the data space. We evaluate our workflow on two samples of the longitudinal cohort study for each of two target variables, respectively.
- Chapter 5 (*Visual Identification of Informative Features*) describes our approach to identify distinct tinnitus phenotypes with parameter-free clustering, and presents novel visualizations to juxtapose phenotypes in a high-dimensional feature space and to explore phenotype-specific characteristics.
- Chapter 6 (*Constructing Evolution Features to Capture Study Participant Change over Time*) present a framework for cohort analysis in longitudinal cohort studies which constructs "evolution features" from latent temporal information describing the change of cohort participants over time. We show that exploiting these novel features improves the generalization performance of classification models and report on results for the longitudinal cohort study.

- Chapter 7 (*Feature Extraction From Short Temporal Sequences for Clustering*) present an approach to build representations from short temporal sequences via clustering by the example of pressure- and posture-dependent plantar temperature and pressure in patients with diabetic foot syndrome.
- Chapter 8 (*Post-Hoc Interpretation of Classification Models*) focuses on making already learned, complex classification models understandable to the domain experts. We provide a workflow that combines classification of high-dimensional medical data and model explanation using post-hoc interpretation methods. To this end, we use Shapely value explanations (SHAP), LASSO coefficients, and partial dependency graphs. Our approach delivers statistics and visualizations representing global feature importance, instance-individual feature importance, and subpopulation-specific feature importance, all of which help illuminate complex black-box machine learning models. We report our results on three applications: (i) tinnitus-related distress in tinnitus patients, (ii) depressivity in tinnitus patients, and (iii) rupture risk in intracranial aneurysms.
- Chapter 9 (*Subpopulation-Specific Learning and Post-Hoc Model Interpretation*) describes an approach that examines how subpopulations differ with respect to their most predictive characteristics in temporal data. To do so, we derive a post-hoc interpretation measure to assess the difference in association of predictors between two subpopulations. We report results for CHA on gender differences (subpopulations of female and male patients) for the two outcomes of tinnitus-related distress and depression, and the effect of treatment for both outcomes.
- Chapter 10 (*Conclusion and Future Work*) concludes the thesis by giving a summary of the contributions and detailed perspectives for the presented work.

Chapter 2

Medical Background & Datasets

Outline

- Types of medical studies
- Short introduction to the medical applications

2.1 Fundamental terms and study types

Typical goals of medical research include identifying long-term determinants and protective factors for an outcome of interest, discovering sub-populations with increased disease prevalence, studying intervention effects by generating statistical (causal) models that explain cause-effect relationships.

Traditional data analysis pipelines are usually structured as follows:

(1) A medical scholar formulates a hypothesis based on observations in clinical practice or current research. Examples: how does a risk factor like alcohol abuse affects the prevalence of a certain outcome? Which effect does a novel therapy have on patients with depressive symptoms?
(2) To investigates this hypothesis, a small set of relevant variables is chosen, which might be controlled for confounders. Adequate data are collected. Variable selection may incorporate controlling for confounders.
(3) The strength of the associations between the selected variables and the outcome is assessed with regression models and statistical methods.
(4) Based on the results, inferential statistics are computed and conclusions are drawn which may foster the implementation of novel preventive measures or the application of suited treatment forms to high-risk patients.

Primary medical research can be classified into basic, clinical and epidemiological studies.

2.2 Basic research

Basic medical research (or experimental research) aims to improve understanding of cellular, molecular and physiological mechanisms of the human health and diseases by conducting cellular and molecular investigations, animal experiments, studies on drug and material properties in strictly controlled laboratory environments [153]. To investigate (causal) effects of one or more variables of interest on the outcome, all other variables are usually kept constant and only those variables of interest are varied. The carefully standardized experimental conditions of basic medical studies ensure a high internal validity, but these conditions often cannot be easily transferred to clinical practice without a loss of generalizability of the results.

!!! Basic research also includes the development and improvement of analytical (e.g., analytical determination of enzymes, markers, genes) and imaging measurement procedures (e.g., computer tomography, magnetic resonance tomography) as well as gene sequencing (e.g., the relationship between eye color and a specific gene sequence), and the development of biometric procedures such as statistical test procedures, modeling and statistical evaluation strategies.

A brief overview of important medical study types is given (see also [175, 153].

2.3 Clinical studies

Clinical studies are generally classified into *interventional* (also: *experimental*) studies and *non-interventional* (also: *observational*) studies.

The general goal of an interventional study is to compare different treatments within a patient population whose members differ as little as possible, except for the treatment branch. A common example is a pharmaceutical study which aims to validate the efficacy and harmlessness of a drug by investigating or establishing main and adverse effects, resorption, metabolization and excretion of the drug. Selection bias can be avoided by appropriate measures, in particular by randomly assigning patients to the groups.

The treatment can be a drug, a surgical procedure or the therapeutic use of a medical product (e.g. stent), but also physiotherapy, acupuncture, psychosocial intervention, rehabilitation, training or diet.

A *randomized controlled trial* (RCT) is considered as study design "gold-standard" as it minimizes selection bias a. by randomly allocating patients to treatment and control group, and b.by ensuring equal distribution of known and unknown influencing variables (confounders), such as risk factors, comorbidities and genetic variabilities. Thus, RCTs are suitable for obtaining an unambiguous answer to a clear question and proving causality.

In contrast, non-interventional clinical studies are patient-related observational studies in which patients receive an individually defined therapy (or all patients receive the exact same therapy). Data analysis often takes place retrospectively. Observational studies are often used to generate hypotheses. - example: study that investigates the regular use of drugs in therapies the treatment, including diagnosis and monitoring, does not follow a pre-defined trial protocol, but only medical practice

2.3.0.1 prospective

Prospective studies can be recognized by the chronological sequence of hypothesis generation and data collection. First, the hypotheses to be tested are determined, for example with respect

to a new treatment procedure. Then, data are collected specifically for hypothesis testing. By formulating testable hypotheses first, it can be ensured that the research questions can actually be answered with the measured data.

2.3.0.2 retrospective

data collection has already taken place before the start of the study

examples:

digitalized medical databases

2.4 Epidemiological studies

TODO: Siehe Buch

Epidemiological studies are usually interested in the distribution and temporal change of the frequency of diseases and their causes in the general population or in sub-populations. Analogous to clinical studies, epidemiology also distinguishes between experimental and observational studies. ???

A *cross-sectional* study (or prevalence study) is conducted only once whereas in a *longitudinal* study, the same study is carried out at several points in time and the results of the individual *waves* of investigation are compared. Longitudinal studies are further categorized into *trend* and *panel* design. In a *trend* study, each wave can involve a different participant sample, i.e., an individual participant is not followed over time. In contrast, a *panel* study investigates the same population at multiple points in time which allows to also measure *intra-individual* temporal changes.

2.4.0.1 controlled vs. uncontrolled

- epidemiological vs clinical vs experimental study
- what is a cohort
- samples (clinical samples, medical research samples) -> validity?

2.5 Challenges

- heterogeneity (numeric, categorical)
- pre-processing of raw image data (non-tabular data)
- different requirements for different studies

2.6 Medical Applications

2.6.1 The Study of Health in Pomerania (SHIP)

After the reunification of Germany, it was found that life expectancy was considerably lower in the east than in the west [190]. Furthermore, there were regional differences within the former East Germany, with the lowest life expectancy in the northeast [190, 194]. To investigate the

causal relationship between the high mortality in the northeastern German population and its risk factors, the research center for community medicine established the *Study of Health in Pomerania* (SHIP) [188], a longitudinal epidemiological study of two independent cohorts in the northeast of Germany. SHIP attempts to describe a wide spectrum of health-related conditions rather than focusing on a specific target disease [188]. In particular, major study goals comprise investigations regarding prevalence of common diseases and their risk factors, correlation and interaction between risk factors and diseases, progression from subclinical to manifest diseases, identification of subgroups with elevated health risk, prediction of incidental diseases, as well as utilization and costs of medical devices. TODO: shorten

Cohort inclusion criteria were age from 20 to 79 years, main residency in the study region and German nationality. Participants of SHIP underwent an extensive, recurring (ca. every 5-6 years) examination program that encompasses personal interviews, body measurements, exercise electrocardiogram, laboratory analysis, ultrasound examinations and full body magnetic resonance tomography (MRT). Figure @ref:(fig:ship-sankey-plot) illustrates participant response and age distribution of show-ups across study waves. Baseline examinations for the first cohort were performed between 1997 and 2001 (SHIP-0, N=4308). Followup examinations were done in 2002-2006 (SHIP-1, n=3300), 2008-2012 (SHIP-2, n=2333), 2014-2016 (SHIP-3, n=1718) and since 2019 (SHIP-4). Baseline information for a second, independent cohort (SHIP-Trend-0, N=4420) was collected between 2008 and 2012 and a followup was conducted between 2016 and 2019. Major strengths of SHIP are a high level of quality assurance, standardized examination protocols, and a high cohort representativeness.

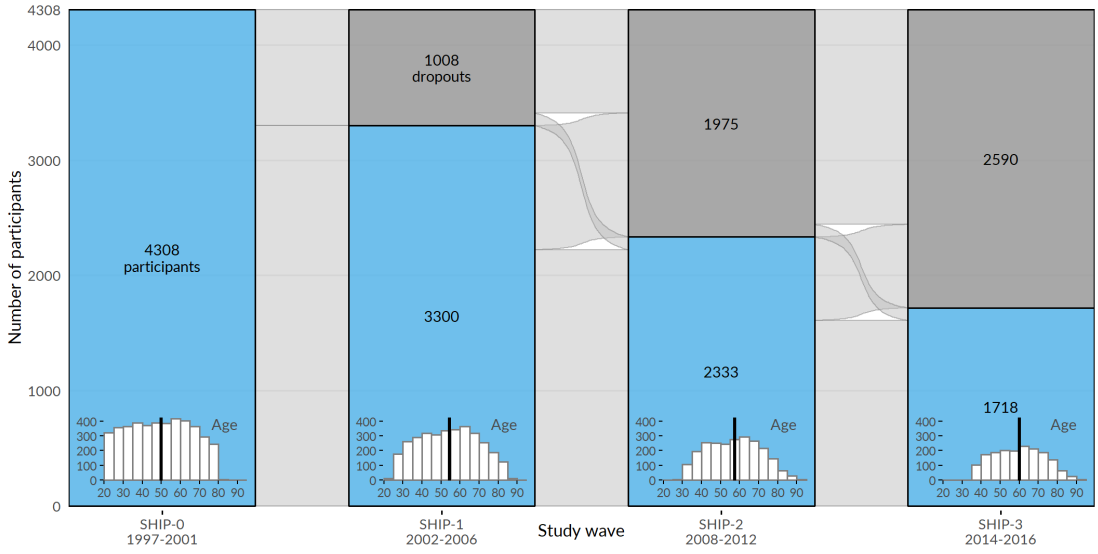

Figure 2.1: **Participation response and age distribution of show-ups across study waves in SHIP.** Average population age is shown as black vertical line in the histograms.

The examination program changed across waves. For example, magnetic resonance imaging (MRI) was only conducted since SHIP-2; liver ultrasound was carried out in SHIP-0 and SHIP-2, but not in SHIP-1; dermatological examinations were conducted in SHIP-1 and SHIP-2, but not in SHIP-0.

In our analyses, we focus on the disorder hepatic steatosis, also known as fatty liver, a condition which is characterized by a high accumulation of fat in the liver and present in ca. 30% of all adults [188, 189]. Risk factors include alcohol abuse, obesity, metabolic syndrome, diabetes, and hyperlipidemia [5]. A liver biopsy is considered as diagnosis gold standard [5], but is

associated with moderate risk for the patient. Non-invasive diagnosis forms include MRT, CT and ultrasound. Since hepatic steatosis is mostly asymptomatic, it often remains undiscovered which may develop into a more serious disease, such as steatohepatitis, cirrhosis, liver cell carcinoma or liver failure.

2.6.2 The Diabetic Foot Clinical Trial (DF)

–>

Diabetic foot syndrom is an umbrella term of foot-related problems in diabetes patients. Up to one out of four diabetes patients develop a foot ulcer during their lifetime [19] with many of them facing amputations in the next four years [111]. More than 85% of foot amputations relate to foot ulcers [65, 185]. The rate of foot amputations among diabetic patients has been estimated to be 17–40 times higher than in the general population [45]. DFS patients are predisposed to a peripheral sensoric neuropathy, which, for example, have the consequence that patients are unaware of the temperature of their feet, or the pressure they apply on them. Affected individuals may even injure themselves without noticing. Excessive plantar pressure loads may aggravate tissue destruction, increasing the lifetime risk of a foot ulceration [166]. However, the understanding of the pathomechanisms underlying tissue destruction without trauma is limited.

At the Medical Faculty of the Otto von Guericke University Magdeburg, an experimental study with 31 healthy volunteers and 30 diabetes patients diagnosed with severe polyneuropathy was conducted to quantify pressure- and posture-dependent changes of plantar temperatures as a surrogate of tissue perfusion. For this purpose, plantar pressure and temperature changes in the feet were recorded during extended phases of standing. Custom-made shoe insoles equipped with eight temperature sensors and eight pressure sensors at preselected positions were used for data acquisition (2.2 (a)). The insoles were positioned into closed protective shoes specifically developed for diabetes patients. Within such shoes the temperature increases over time due to exchange with the body temperature of the user and is also affected by the environmental temperature. To closely monitor the in-shoe temperature changes one sensor was placed at the bottom of the insole without contact to the feet, which was denoted “ambient temperature sensor”.

The data acquisition started immediately after putting on the shoes. The participants were asked to follow a predefined sequence of actions, that are alternating postures of standing (stance phase) and seating (pause). The sessions consisted of six stance episodes, lasting 5, 10, 20, 5, 10 and 20 minutes each, separated by seating episodes lasting 5 min each. The participants were instructed to apply pressure equally on both feet while standing. Participants did not receive immediate feedback of the actual pressure application during the sessions, however study nurses verbally encouraged to keep the pressure without release while standing. In the seated position the participants were instructed to release pressure for 5 minutes, while still keeping contact to the insole. The participants were explicitly asked to adhere to these instructions, e.g. they should not temporarily release pressure during a stance episode. The study protocol furthermore encompassed that the measurements were performed twice, once at room temperature of ca 22°C and once outdoors with an ambient temperature of ca. 16°C. The two measurements were performed on two independent days.

The termographic images in Figure 2.2 (b) exemplarily visualize changes of plantar temperature, for a healthy volunteer in a seated position, before pressure application (1), after placement of a 20 kg weight on the front of the upper thighs (2-6), and after removal of the additional weight (7-8). During pressure load, a gradual temporal temperature decrease was detected predominantly in the forefoot. After pressure release, a rapid temperature rise within 1 min was observed.

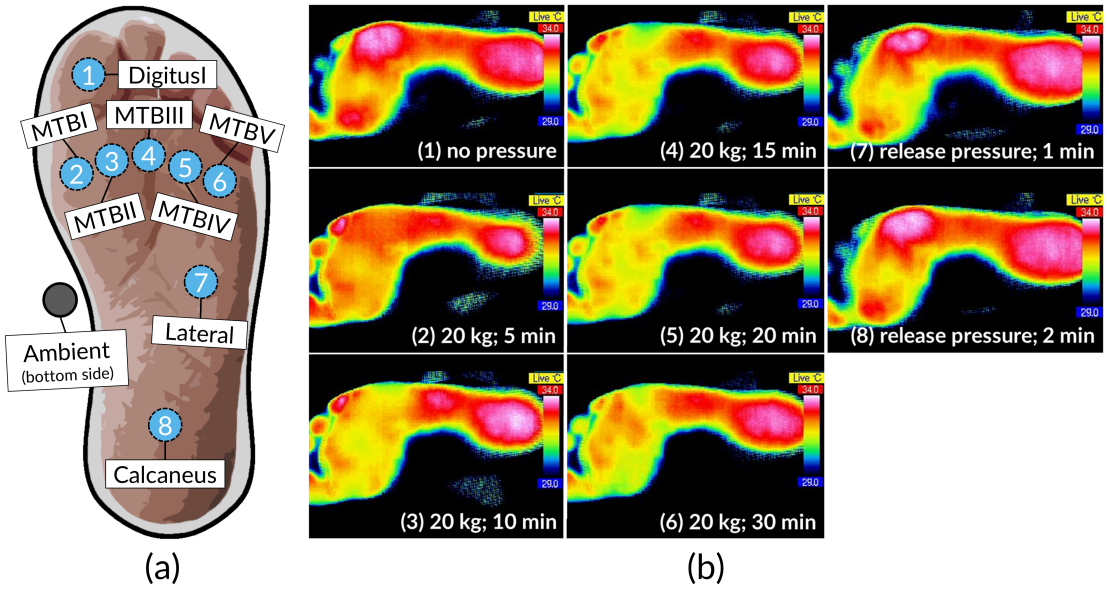

Figure 2.2: (a) Sensor positioning in relation to foot placement. (b) Infrared images of a healthy volunteer in a seated position without pressure application to the feet (before) and following positioning of a 20 kg weight on both upper thighs. A time-dependent decrease of temperature was detected predominantly in the forefoot, visualized by yellow color during the pressure load. Within 1 min of pressure release, a rapid temperature rise was detected.

2.6.3 The Intracranial Aneurysm Angiography Image Dataset (IAD)

Retrospective clinical

Intracranial aneurysms are pathologic dilations of the intracranial vessel wall, often in form of a dilation. They bear a a risk of rupture which then lead to subarachnoidal hemorrhages with often fatal consequences for the patient. Since treatment can also cause severe complications, extensive studies were conducted to assess the patient-individual rupture risk based on various parameters, including aneurysm symptomatology, size and location, as well as patient age and gender [192]. Further studies identified parameters, such as aspect ratio, undulation index and nonsphericity index as statistically significant with respect to aneurysm rupture status [37, 198]. However, although these studies allow for a retrospective analysis, the clinician needs further guidance in case an asymptomatic aneurysm (as accidental finding) was detected and the rupture risk should be determined.

We developed methods for the *"Intracranial Aneurysm Angiography Image Dataset"* (IAD) comprising 3D rotational angiography data from 74 patients (age: 33-85 years, 17 male and 57 female patients) of the university hospital of Magdeburg, Germany, adding up to a total of 100 intracranial aneurysms. We identified two primary goals for this dataset: (i) build models that can accurately predict rupture status based on morphological parameters only, and (ii) assess the importance of these parameters to the models with optimal accuracy.

Inspired by the results of Baharoglu et al. [12] who found differences between sidewall and bifurcation aneurysms (cf. Figure 2.3) regarding the relationship of several morphological parameters and rupture status, we learn distinct models for the subset of sidewall aneurysms (9 (37.5%) of 24 ruptured) and for the subset of bifurcation aneurysms (29 (46.8%) of 62 ruptured). Moreover,

we run experiments on a combined group (43 of 100 ruptures) which includes 14 samples that could not be clearly determined to be either sidewall or bifurcation aneurysms.

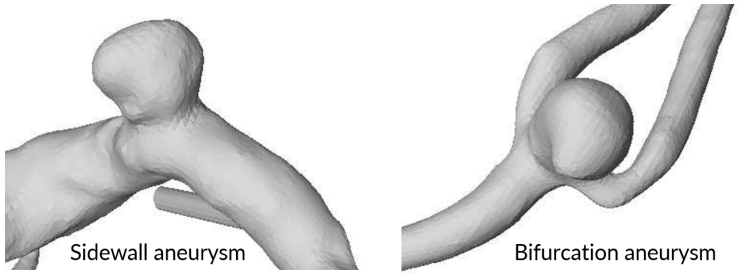

Figure 2.3: **Sidewall and bifurcation aneurysm.** Illustration of a sidewall aneurysm at the side of the parent vessel wall (left) and a bifurcation aneurysm at a vessel bifurcation (right).

2.6.4 The Tinnitus Patients Observational Therapy Study Dataset (CHA)

–> –>

Tinnitus is the perception of a phantom sound in absence of an external sound source. It is a complex multi-factorially caused and maintained phenomenon, and estimated to affect between 10% and 15% of the adult population [11]. The associated annual economic burden amounts to US$19.4 billion in the United States [16] and €6.8 billion in the Netherlands alone [128]. Clinical assessment of tinnitus is challenging due patient heterogeneities with respect to perception of tinnitus (laterality, pitch, sound characteristics, frequency, permanence, chronicity), risk factors (including hearing loss, temporomandibular joint disorder, aging), comorbidities (including hyperacusis, depression, sleep disorders), perceived distress, and treatment response [27]. These differences make the identification of a suitable and effective form of treatment difficult. Currently, there does not exist a therapy gold standard: sound therapy (masking), informational counseling (minimal contact education), cognitive behavioral therapy and tinnitus retraining were found to be effective for some patients, but there is also evidence that not all patients benefit equally from these treatment forms [87, 106, 78, 130, 141].

Due to the heterogeneous nature of the tinnitus symptom as well as the unclear evidence-base as to its treatment and management, the identification of patient subgroups is vital to stratify individual pathophysiology and treatment pathways [113, 178, 112].

The “tinnitus patients observational therapy study dataset” (*CHA*) comprises self-report data from 4,103 tinnitus patients who had been treated at the Tinnitus Center of Charité University Medicine Berlin between January 2011 and October 2015. All patients were 18 years of age or older and had been suffering from tinnitus for at least 3 months. Exclusion criteria were presence of acute psychotic illness or addiction disorder, deafness and insufficient knowledge of the German language. Treatment comprised a multimodal 7-day program that included intensive and daily informational counseling, detailed ear-nose-throat as well as psychological diagnostics, cognitive behavior therapy interventions, hearing exercises, progressive muscle relaxation and physiotherapy. At baseline (*T0*; before therapy commencement) and after treatment (*T1*), patients were asked to complete multiple self-report questionnaires. These questionnaires were selected to obtain a comprehensive tinnitus assessment, including tinnitus-related distress and the psychosomatic background of tinnitus with anxiety, depression, general quality of life and experienced physical impairments. Table 2.1 provides an overview of all questionnaires used in

the analysis. Most questionnaires contain multiple-choice items with answers on a Likert scale. For example, the TQ contains 52 statements, such as "I am unable to enjoy listening to music because of the noises.", and respondents can give 3 possible answers: "not true" (encoded as 0), "partly true" (1) and "true" (2). Some questionnaires also comprise aggregated variables, called "subscales" and "total scores". For example, the TQ total score (TQ_distress) is calculated as sum of 40 item values, where 2 items are used twice [61], yielding a value range from 0 to 84, where higher values represent higher tinnitus-related distress. The cutoff value 46 [61] is used to distinguish between *compensated* (0-46) and *decompensated* (47-84) tinnitus. Further, for each questionnaire, the average time to answer an item was recorded. Figure @ref:(fig:02-cha-patient-demographics-plot) provides a graphical illustration of demographics for 3,803 (92.7%) patients with complete data for the SOZ questionnaire and TQ score.

Table 2.1: Description of questionnaires that form the basis for CHA.

	Name	**Scope**	**Scales**	**|F|**
1	ACSA: Anamnestic Comparative Self-Assessment [15]	Quality of life	–	1
2	ADSL: General Depression Scale [149, 76]	Depressive symptoms	–	22
3	BI: Berlin Complaint Inventory [90]	General well-being, autonomic nervous system, pain and emotionality	Exhaustion, abdominal symptoms, limb pain, heart symptoms	29
4	BSF: Berlin Mood Questionnaire [89]	Mood	Anger, anxious depression, apathy, elevated mood, fatigue and mindset	36
5	ISR: ICD-10 Symptom Rating [177]	Mental disorders	Anxiety, depression, obsessive-compulsive syndrome, somatoform syndrome, eating disorder, additional items	36
6	PHQK: (Short-form) Patient Health Questionnaire [167]	Symptoms of depression and anxiety	–	17
7	PSQ: Perceived Stress Questionnaire [49]	Stress	Demand, tension, joy, worries	35
8	SES: Pain Perception Scale [57]	Pain	Affective pain, sensoric pain	26
9	SSKAL: Visual Analogue Scales Pain	Pain impairment, frequency and intensity	–	3
10	SF8: Short Form 8 Health Survey [23]	Health-related quality of life	Bodily health, overall health, mental health, physical functioning, role emotional, role physical, vitality, mental component, physical component	18
11	SOZK: A socio-demographics questionnaire [22]	Gender, partnership status, education, employment status, among others	–	27

	Name	Scope	Scales	|F|
12	SWOP: Self-Efficacy-Optimism-Pessimism Scale questionnaire [160]	Self efficacy, optimism, pessimism	–	12
13	TINSKAL: Visual analogue scales	Tinnitus loudness, frequency and distress	–	3
14	TLQ: Tinnitus Localization and Quality questionnaire [60]	Location (left, right, bilateral, entire head) and sound of tinnitus	–	8
15	TQ: Tinnitus Questionnaire (German version) [61]	Tinnitus-related distress and tinnitus severity	Emotional and cognitive burden, persistence of sound, hearing difficulties, sleep difficulties and somatic complaints	60

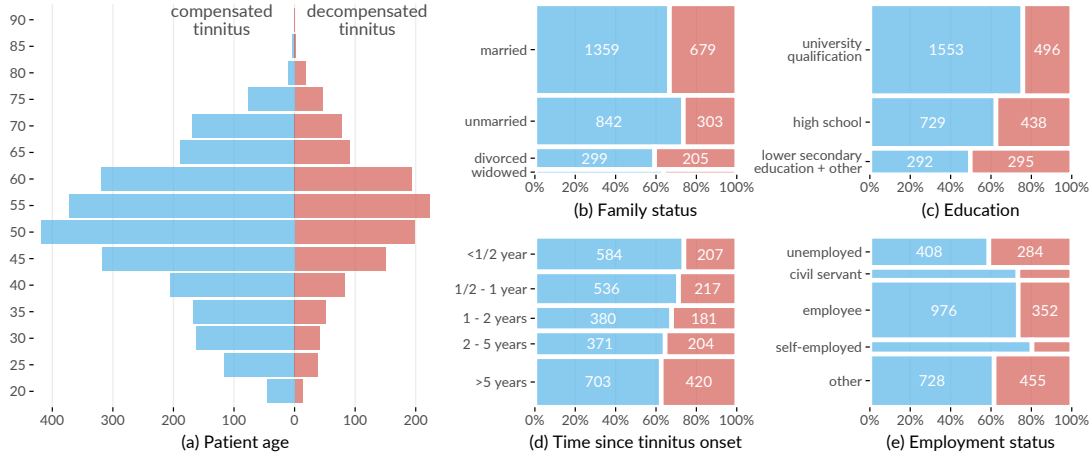

Figure 2.4: **Patient demographics (CHA).** Overview of patient demographics by degree of tinnitus distress measured before therapy commencement.

–> –> –> –>

Part I

SUBPOPULATION DISCOVERY IN HIGH-DIMENSIONAL DATA

Chapter 3

Interactive Discovery and Inspection of Subpopulations

Brief Chapter Summary

We study the separation between two positive and a negative outcome for a disease in epidemiological data. Our goal is to identify compact subpopulations that are as pure as possible with respect to the target variable.

This chapter is partly based on:
Uli Niemann, Henry Völzke, Jens-Peter Kühn, and Myra Spiliopoulou. "Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis". In: *Expert Systems with Applications* 41.11 (2014), pp. 5405-5415. DOI: 10.1016/j.eswa.2014.02.040.

3.1 Motivation and Comparison to Related Work

Medical decisions on diagnosis and treatment of multifactorial conditions such as diseases and disorders are based on clinical and epidemiological studies; the latter accommodate information on participants with and without a condition and allow for learning discriminative models and, in the longitudinal design, for understanding the progress of the condition. For example, several studies identified risk factors (like obesity or alcohol consumption) and co-morbidities (like cardiovascular diseases) associated with hepatic steatosis ("fatty liver") [94, 117, 168, 173, 129]. However, these studies identified risk factors and associated outcomes that pertain to the whole population. Our work originated from the necessity to identify such factors and outcomes for subpopulations and thus to promote personalized diagnosis and treatment, as expected in personalized medicine [85, 186].

Classification on subpopulations was studied by Zhanga and Kodell [201] who pointed out that classifier performance on the whole dataset can be low if the complete population is very heterogeneous. Therefore, they first trained an ensemble of classifiers, and then used the predictions of each ensemble member to create a new feature space. They performed hierarchical clustering

to partitioned the instances into three subpopulations: one where the prediction accuracy is high, one where it is intermediate and one where it is low. With this approach, Zhanga and Kodell split the original dataset into subpopulations that are easy or difficult to classify. While the method seems appealing in general, it appeared unsuitable for the three-class problem of the SHIP data which exhibits a very skewed distribution, hence it is clear that low classification accuracy is (partially) caused by the skew. Hence, we studied the dataset exploratively *before* classification, to identify less skewed subpopulations, and exploratively *after* classification, to identify - inside each subpopulation - variables which are highly associated to the outcome.

Pinheiro et al. performed association rule discovery on patients with liver carcinoma [142]. The authors pointed out that early detection of liver cancer may help reducing the five-year mortality rate, but early detection is difficult, because in the onset of a liver carcinoma, the patient often observes no symptoms [142]. Pinheiro et al. leveraged the association rule algorithm FP-growth [73] to discover high-confidence association rules and high-confidence classification rules regarding mortality in liver cancer patients. We also considered association rules promising for the analysis of medical data, because they are easy to compute and deliver results that are understandable by humans. Therefore, we also used association rules as baseline method, though for epidemiological data and for classification rather than mortality prediction. To use association rules for classification, we specified that the rule consequent is the target variable.

Zhang et al. [200] addressed the increasing technical challenges of medical expert-driven subpopulation discovery due to increasingly large and complex medical data which often comprise information of hundrets of variables for thousands of patients in form of tables, images or text. Whereas in the past it was sufficient for a physician to have working knowledge of basic statistics and a spreadsheet software like Microsoft Excel to analyze a small table with patient data, nowadays more effective and efficient approaches are required for management, analysis and summarization of very large medical data.
As a result, domain experts usually rely on technical experts to help them perform these tasks. However, this back and forth process is often slow, tedious and expensive. Hence, it would be better to equip the domain expert with a technical tool that allows them to quickly perform exploratory analyses on their own. Zhang et al. [200] presented CAVA, a system that incorporates miscellaneous subgroup visualizations (called "Views") and analytic components (called "Analytics") for subgroup comparison. The main panel in Figure 3.1 shows one of the Views: a flow diagram [196] of patient subgroups with the same sequence of symptoms. Users can obtain additional summaries by interacting with the visualization, e.g. by placing one of the boxes in the flow diagram onto one of the entries in the Analytics panel via drag-and-drop. Further, user can expand the selected cohort by letting the tool search patients who do not strictly match the current inclusion criteria but are somewhat *similar* to the current patient subpopulation [42].

Krause et al. [105] argued that model selection should not be based solely on global performance metrics like accuracy, because these statistics do not contribute to a better understanding of the model's reasoning. Furthermore, a complex yet very accurate model does not automatically warrant actionable insights. Krause et al. propose Prospector [105], a system that provides diagnostic components for complex classification models based on the concepts of partial dependence (PD) plots [53]. PD plots are a popular tool to visualize the marginal effect of a feature on the predicted outcome probability. Briefly, each point of a PD curve represents the average prediction of the model over all observations, given that these observations had a fixed value for a feature of interest. A feature whose PD curve has a high range or variability is considered to be more impactful to the model prediction than a feature with a flat PD curve. Closely related to PD plots are individual conditional expectation (ICE) plots [62], which display one curve for every observation, and thus help to uncover contrasting subpopulations who might "average out" in a PD plot. Prospector combines PD and ICE curves to depict the relationship between a feature and the model prediction on a (*global*) model level and a (*local*) patient-individual

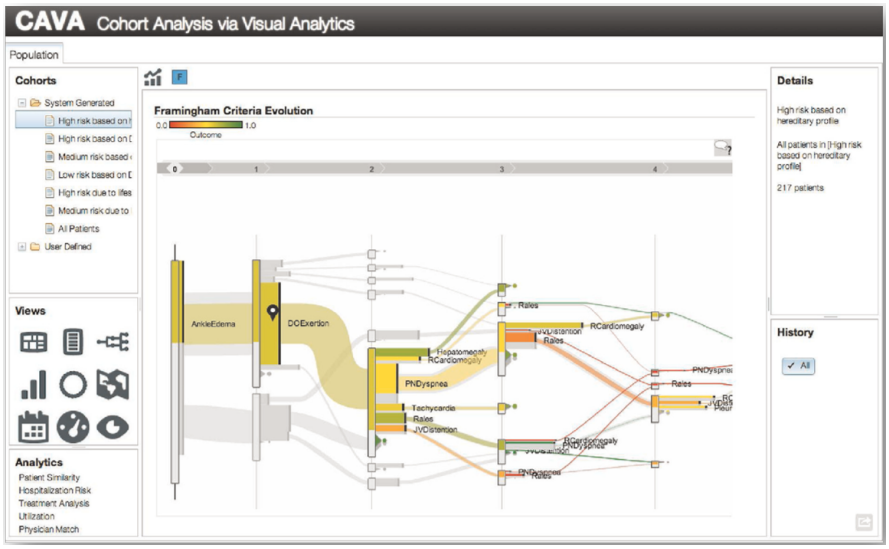

Figure 3.1: CAVA's graphical user interface. The flow chart visualizes cardiac patient subgroups organized by shared symptom occurrence. Color represents hospitalization risk. The user can switch between graphical representations and data processing methods via drag-and-drop operations. The top-right panel provides detailed information for the currently selected patients. The bottom-right panel contains a provenance graph that allows the user to undo operations and to revisit previous interaction steps. (from: [200])

level. Further, a color bar is provided as more compact alternative to ICE curves (Figure 3.2) (a). A stacked bar chart shows the distribution of predicted risk scores for each study group (Figure 3.2 (b)), and the user can click on a specific decile to get a list of individual patients with their exact prediciton score and label. In this way, patients whose prediction score is close to the decision boundary can be further investigated. The authors calculate for each feature the "most impactful feature change": given the actual feature value of a patient, they identify a near counterfactual value that led to a large change in the predicted risk score, by minimizing difference to the original feature value and maximizing predicted risk score. The top-5 of these so called "suggested changes" are displayed – separately for inreasing and decreasing disease risk – in a table (cf. Figure 3.2 (c)), and incorporated as interactive elements into the IC color bars (cf. Figure 3.2 (d)).

Pahins et al. [138] presented COVIZ, a system for cohort construction in large spatiotemporal datasets. COVIZ comprises mechanisms for explorative data analysis of treatment pathways and event trajectories, visual cohort comparison and visual querying. One of COVIZ' design goals is being fast, e.g. by using efficient data structures like Quantile Data Structure [116] to ensure low latency for all computational operations and hence suitablity for large datasets.

Bernard et al. [14] proposed a system for cohort construction in temporal prostate cancer cohort data which comprised visualizations for both subgroups and individual patients. To guide users during exploration, visual markers indicate interesting relationships between attributes derived from statistical tests.

3.1.1 Other related work

- IIComPath: [34]

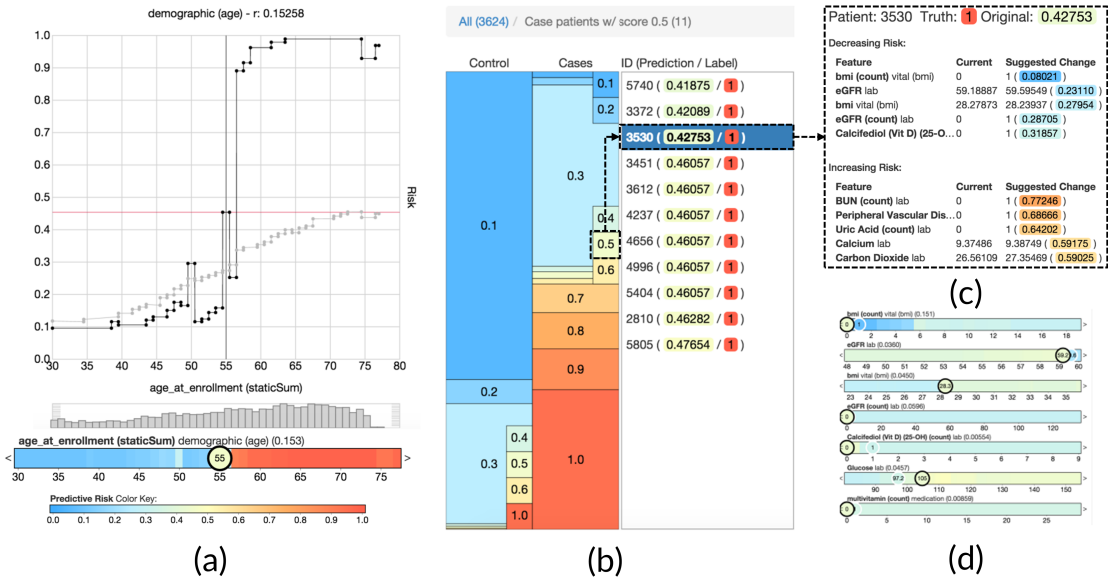

Figure 3.2: A selection of Prospector's model diagnostics. (a) The upper chart depicts two curves for the feature "age": the gray partial dependence (PD) curve represents the marginal prediction of the model over all patients whereas the black individual conditional expectation (ICE) curve illustrates the effect of counterfactual age values on the predicted diabetes risk for an example patient. The histogram shows the age distribution. The color bar placed beneath is a compact depiction of the ICE curve above; the encircled value represents the feature value of the selected patient. (b) Stacked bars show the distribution of predicted risk scores for each study group. Clicking on one of the bars opens a table that shows ID, predicted risk and true label for all patients belonging to the selected prediction risk decile. (c) Summary table of the most "impactful feature changes" for decreasing (upper group) and increasing (lower group) predicted risk: each row shows the actual feature value and the "suggested change", i.e., a similar but counterfactual value that led to a considerable change in predicted risk. (d) Multiple PD color bars augmented with suggested changes (white encircled labels). (adapted from: [105])

3.1.2 Previous Work on SHIP

Klemm et al. [102] presented a "3D regression cube" system which enables interactive exploration of feature correlations in epidemiological datasets. The system generates a large number of multiple regression models from various combinations of one dependent and two independent variables and displays their goodness of fit in a three-dimensional heatmap. The system allows the user to modify the regression equation, for example, by changing the number of independent variables, by specifying interaction terms or by fixing one of the variables to reduce computational complexity or to specifically focus on a variable of interest. Our approach is also capable of identifying variables that are highly associated with the outcome, but we search for subpopulation-specific relationships instead of generating a global model for the whole dataset, and we further provide predictive value ranges.

Klemm et al. [103] presented a system which combines visual representations of non-image and image data. They identify clusters of backpain patients on the SHIP data. As we fixed hepatic steatosis as outcome, we rather opted to build supervised models and classification rules that directly captured the relationships between predictive variables and the outcome.

Alemzadeh et al. [2] presented S-ADVIsED, a system for interactive exploration of subspace clusters, incorporating various visualization types, such as donut charts, correlation heatmaps, scatterplot matrices, mosaic charts and error bar graphs. While S-ADVIsED requires the user to input the mining results obtained in advance outside of the system, our tool allows for an expert-driven interactive subpopulation discovery.

Hielscher et al. [82] developed a semi-supervised constrained-based subspace clustering algorithm to find diverse sets of *interesting* feature subsets using the SHIP data. To guide the search for interesting feature subsets, the expert can provide their domain knowledge in form of instance-level constraints, thus forcing pairs of instances to be assigned either to the same or a different cluster. Hielscher et al. [80] extended their work and introduced a mechanism to compare subpopulations between independent cohorts.

As a consequence, thesame group developed a constrained-based technique, where theclustering is guided by a small set of constraints, given by an expert[HNP*18]. As an example, for a few pairs of participants diagnosedwith fatty liver, the expert specifies that these participants must be inthe same clusters, whereas some other pairs of participants are forcedto be in different clusters since one in each pair is diagnosed withthe disorder and the other is not. This semi-supervised subspaceclustering turned out to yield relevant results for epidemiologists[HNP*18].

TODO: Preim: Visual analytics of image-centric cohort studies in epidemiology

3.2 Subpopulation Discovery Workflow and Interactive Mining Assistant

Our subpopulation discovery workflow is presented hereafter. The dataset used for population partitioning and class sepration on the target variable hepatic steatosis come from the Study of Health in Pomerania (SHIP) which is described in Subsection 2.6.1.

In subsection 3.2.1, explains origin and availability of the target variable. In subsection 3.2.2, the motivation for partitioning of the data and the partitioning steps are presented. Then, the used methods for class separation on the whole dataset and on the partitions are discussed in Subsection 3.2.3.

3.2.1 Target variable

The target variable is derived from the participant's liver fat concentration computed with magnetic resonance imaging (MRI). At the time of writing the original manuscript, MRI results were available only for 578 SHIP-2 participants. We use the data of these participants for classifier learning, while our interactive ruleInspector (cf. section **???**) also juxtaposes these data to the data of the remaining 1755 participants, for whom the MRI recordings were not made available.

Participants with a liver fat concentration of 10% or less are mapped to class A ("negative" class, i.e., absence of the disorder), values greater than 10% and lower than 25% are mapped to class B (increased liver fat / fatty liver tendency), and values greater than 25% are mapped to class C (high liver fat). We consider classes B and C as "positive". Although the cut-off value of 10% is higher than the value of 5% suggested by [110] for separation between subjects with and without hepatic steatosis. However, the primary interest from the medical perspective was the identification of important variables for individuals that are likely to be ill. The selection of a high cut-off value exacerbated class imbalance and made the data analysis more challenging. Figure 3.3 depicts the class distribution by gender. Out of the 578 participants, 438 belong to class A (ca. 76%), 108 to B (ca. 19%) and 32 to C (ca. 6%). Men are more likely to exhibit increased or high liver fat than women (30.7% vs. 18.8% in classes B or C).

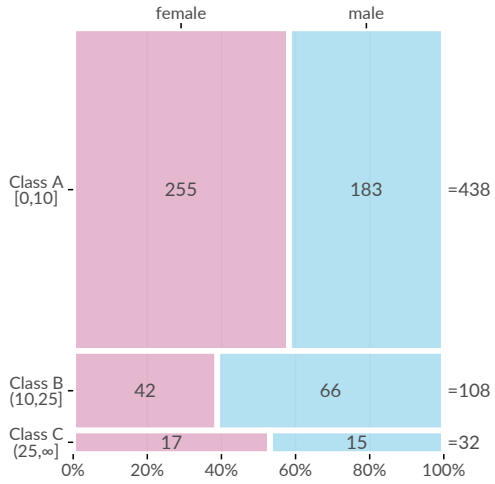

Figure 3.3: Gender-specific distribution of the target variable.

Next to the target variable, the dataset contains 66 variables extracted from participants' questionnaire answers and medical tests (cf. [187]). They are variables on sociodemographics (gender, age, etc.), variables on consumption behaviour (e.g. alcohol and cigarettes), SNPs (genetic information), variables extracted from laboratory data (e.g. sera concentrations), and two variables on the results of the liver ultrasound – `stea_s2` and `stea_alt75_s2`. Both variables take symbolic values that reflect the likelihood that the participant has fatty liver; the latter is a combination of the former and the ALAT recording for the participant; details are in (cf. [187]). Almost all variables mentioned hereafter have the suffix `_s2` which indicates measurements of the SHIP-2 followup, as opposed to SHIP-0 (`_s0`) and SHIP-1 (`_s1`). Exceptions are gender, highest school degree and the 10 SNP variables.

3.2.2 Partitioning the Dataset into Subpopulations

Since the dataset is imbalanced with respect to gender (314 women, 264 men), we decided to partition the dataset before classification. First, we investigated the class distributions in the two partitions on gender. We observed that the distributions are very different, most notably with respect to class B (cf. Figure 3.3. As second step, we studied the class distribution by gender and age, whereupon we detected that age is associated with PartitionF but not with PartitionM. Third, we identified a cut-off point for age by introducing a heuristic that identifies the age value which minimizes the standard deviation with respect to the target variable. Afterwards, we performed supervised learning separately on the partitions of male and female participants. Furthermore, we built an additional learner for the subpopulation of older female participants aged above the cut-off point 52.

To understand how age affects the class distribution, we introduced a heuristic that determines the cutoff age value at which `partitionF` splits into two bins, so that the standard deviations of the liver fat concentration in each bin are minimized. Let splitAge denote the cutoff value and $X_y = \{x \in \mathsf{PartitionF} | \text{age of } x \leq \mathsf{splitAge}\}$, $X_z = \{x \in \mathsf{PartitionF} | \text{age of } x > \mathsf{splitAge}\}$ denote the bins. Further, let n be the cardinality of $X_y \cup X_z$ i.e. of PartitionF. Then, we define the Sum of weighted Standard Deviations ($SwSD$) as

$$SwSD\left(X_y, X_z\right) = \frac{|X_y|}{n}\sigma(X_y) + \frac{|X_z|}{n}\sigma(X_z) \tag{3.1}$$

where $|X_i|$ is the cardinality of X_i and $\sigma(X_i)$ the standard deviation of the original liver fat values. Our heuristic selects the splitAge such that $SwSD$ is minimal. For PartitionF, the minimum value was 7.44 at the age of 52, i.e. close to the onset of menopause.

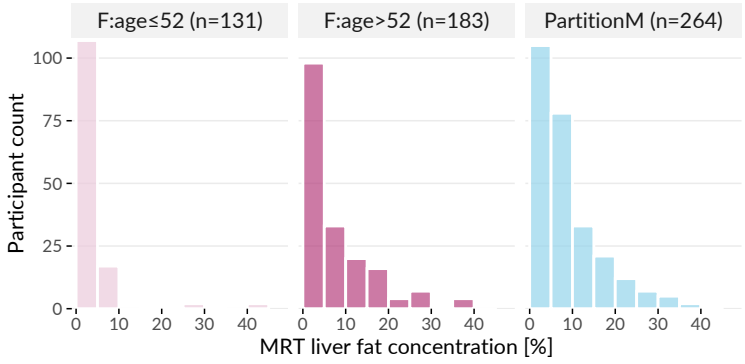

Figure 3.4: Distribution of liver fat concentration in male participants, and in female younger and older than 52 years. The horizontal axis shows the liver fat concentration in bins of 5%, while the vertical axis shows the number of participants in each bin.

The histograms in Figure 3.4 depict the differences in the liver fat concentration distributions at the age cutoff value of 52. Next to PartitionM (n=264), we show the subpartitions $\mathsf{F} : \mathsf{age} \leq 52$ (n=131) and $\mathsf{F} : \mathsf{age} > 52$ (n=183) of PartitionF. Most of female participants in $\mathsf{F} : \mathsf{age} \leq 52$ have no more than 5% liver fat concentration and ca. 95% have no more than 10%, i.e., they belong to the negative class A. In contrast, ca. 28% of the participants in $\mathsf{F} : \mathsf{age} > 52$ have a liver fat concentration of more than 10%; they belong to the positive classes B and C.

For the classification of the cohort participants we concentrated on algorithms that deliver interpretable models, since we wanted to identify predictive *conditions*, i.e., variables and val-

Table 3.1: **Cost matrix.** Cost matrix penalizing misclassification under class skew.

		Predicted class		
		A	B	C
True class	A	0	1	2
	B	2	0	1
	C	3	2	0

ues/ranges in the models. Hence, we considered decision trees, classification rules and regression trees.

We employed the J4.8 decision tree classification algorithm (equivalent to the C4.5 algorithm [145]) of the Waikato Environment for Knowledge Analysis (Weka) workbench [51]. This algorithm builds a tree successively, by splitting each node (subset of the dataset) on the variable that maximizes information gain within that node. The original algorithm operates only on variables that take categorical values and creates one child node per value. However, the implementation in the Weka library also provides an option that forces the algorithm to always create exactly two child nodes: one for the best separating value and one for all other values. We used this option in our experiments, because it delivers trees of better quality. Moreover, the Weka algorithm also supports variables that take numeric values: a node is split into two child nodes by partitioning the value range of the variable into two intervals.

To deal with the skewed distribution, we considered the following classification variants:

- *Naive*: the problem of imbalanced data is ignored.
- *InfoGain*: we keep only the top-30 of the 66 variables, by sorting the variables on information gain towards the target variable.
- *Oversampling*: We use SMOTE [29] to resample the dataset with minority-oversampling: for class B, 100% new instances are generated, for class C 300% new instances are generated, resulting in following distribution A:438, B:216, C:128.
- *CostMatrix*: We prefered to misclassify a negative case rather than not detecting a positive case, so we penalized false negatives (FN) more than false positives (FP). We used the cost matrix depicted in Table 3.1.

3.2.3 Classification Rule Discovery

Classification rules can uncover interesting relationships between one or more features and the outcome [54, 79]. Compared to model families such as deep neural networks, support vector machines and random forests, classification rules achieve inferior accuracy most of the time. However, they are easier to interpret and reason about, and hence lend themselves better for interactive subpopulation discovery. In epidemiological research, interesting subpopulations could be subsequently used to formulate and validate a small set of hypotheses or just to explore associations between risk factors for a specific outcome. An interesting subpopulation could be phrased as "In the sample of this study, the prevalence of goiter is 32%, while the likelihood in the subpopulation described by *thyroid-stimulating hormone* smaller equal 1.63 mU/l and *body mass index* greater than 32.5 kg/m^2 is 49%".

Classification rule algorithms induce descriptions of "*interesting*'' subpopulations of the data where interestingness is quantified by a quality function. A classification rule is an association

rule whose consequent is fixed to a specific class value. Consider the exemplary classification rule r_1:

$$r_1 : \underbrace{som\_waist\_s2 < 80 \wedge age\_ship\_s2 > 59\,(\wedge ...)}_{\text{Antecedent}} \longrightarrow \underbrace{hepatic\_steatosis = pos}_{\text{Consequent}} \tag{3.2}$$

Underpinnings

Classification rules are expressed in the form of r : antecedent $\longrightarrow T = v$. The conjunction of *conditions* (i.e. feature - feature value pairs) left to the arrow constitutes the rule's antecedent (or left hand site). In the consequent (or right hand side), v is the requested value for the target variable T.

We define $s(r)$ as the subpopulation or *cover set* of r, i.e. the set of instances that satisfy the antecedent of r. The *coverage* of r, which is the fraction of instances covered by r, is then defined as $Cov(r) = |s(r)|/N$, where N is the total number of instances. The *support* of r quantifies the percentage of instances covered by r that additionally have $T = v$, calculated as $Sup(r) = |s(r)_{T=v}|/N$. The *confidence* of r (also referred to as precision or accuracy) is defined as $Conf(r) = |s(r)_{T=v}|/|s(r)|$ and expresses the relative frequency of instances satisfying the complete rule (i.e., both the antecedent and the consequent) among those satisfying only the antecedent. The *recall* or *sensitivity* of r with respect to $T = v$ is defined as $Recall(r) = Sensitivity(r) = \frac{|s(r)_{T=v}|}{n_{T=v}}$. The *Weighted Relative Accuracy* of a rule is an interestingness measure which balances coverage and confidence gain and is often used as internal quality criterion for candidate generation [79]. It is defined as $WRAcc(r) = Cov(r) \cdot (Conf(r) - \frac{n_{T=v}}{N})$.

As an example, Figure 3.5 illustrates an exemplary rule r_2 in a dataset with 10 instances and a binary target, where circles in cyan color represent instances from the negative class and red circles are positive instances. The cover set of r_2 contains instances 7, 8, 9 and 10, hence $Cov(r_2) = 0.40$. Further, $Sup(r_2) = 0.30$, $Conf(r_2) = 0.75$ and $WRAcc(r_2) = 0.4 \cdot (0.75 - 0.4) = 0.14$.

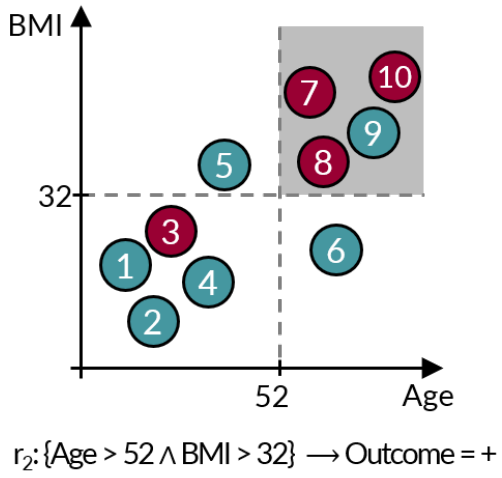

Figure 3.5: Exemplary classification rule.

3.2.4 HotSpot

For classification rule discovery we use the algorithm HotSpot[1] provided in the Waikato Environment for Knowledge Analyis (WEKA) workbench [51]. HotSpot is a beam search algorithm that implements a general-to-specific approach for extracting rules. A single rule is constructed by successively adding the condition to the antecedent that locally maximizes confidence. Contrary to general hill-climbing which considers only the best rule candidate at each iteration, HotSpot's beam search keeps the b highest ranked candidates and refines them in later steps. Consequently, HotSpot reduces the *"myopia"* [54] hill-climbing search typically suffers from. Briefly, hill-climbing approaches consider only the locally optimal candidate at each iteration. As a consequence a globally optimal rule is not found if it is not locally optimal in every iteration. From an application point of view, it is also desirable to generate more than one rule, as alternative descriptions of subpopulations can facilitate hypothesis generation. The beam width can be specified as `maximum branching factor`, the maximum number of conditions that may be added to a rule candidate. In every iteration, rule candidates must satisfy the `minimum value count` which is the sensitivity threshold. To avoid adding a condition that leads only to a marginal improvement in confidence, the parameter `minimum improvement`, i.e. the minimum relative improvement in confidence by adding a further condition can be specified. Computational complexity of the rule search can be reduced by specifying a `maximum rule length`, which is the number of conditions in the antecedent.

describe parametrization in experiments

3.2.5 Interactive Medical Miner

Classification rules can provide valuable insights on potentially prevalent conditions for different subpopulations of the cohort under study. However, if the number of rules produced is large, as it is usually the case in large epidemiological data, the conditions of the rules overlap and some conditions are present under each of the target feature's classes. Hence, the medical expert needs inspection aids to decide which rules are informative and which features should be studied further. Our Interactive Medical Miner (IMM) allows the expert to (a) discover classification rules subject to frequency constraints, inspect the frequency of those rules (b) towards each class and (c) against the unlabeled part of the cohort, and (d) study the statistics of each rule for the values of selected variables. We describe these functionalities below, referring to the screenshot in Figure 3.6.

The user interface consists of six panels. The "Settings" panel (upper left) allows the medical expert to set the parameters for rule induction before pressing the button "Build Rules". Below this panel, the discovered rules are displayed. The panel "Sorting preference" allows the expert to specify whether rules should be sorted by confidence, by coverage, or rather alphabetically for better overview of overlapping rules.

The mining criteria include the dataset (choosing between the whole dataset versus one of the partitions), the class for which rules should be generated (drop-down list "Class") and the constraints with respect to this class, i.e., `min value count` (which can also be given as relative number), `maximum rule length`, `maximum branching factor` and `minimum improvement`. As an example how these parameters affect the rule search, consider the selected rule in Figure xxx, som_huef_s2 > 109 & crea_u_s2 $> 5.38 \longrightarrow$ mrt_liverfat_s2 = B, which has a coverage of 0.12 and a confidence of 0.56. The sensitivity of $38/108 = 0.352$ satisfies the minimum value count threshold of 0.33. From the Apriori property, it is apparent that each of the two conditions in the rule's antecedent, namely som_huef_s2 > 109 and crea_u_s2 > 5.38 must also exceed this

[1] https://weka.sourceforge.io/packageMetaData/hotSpot/1.0.4.html

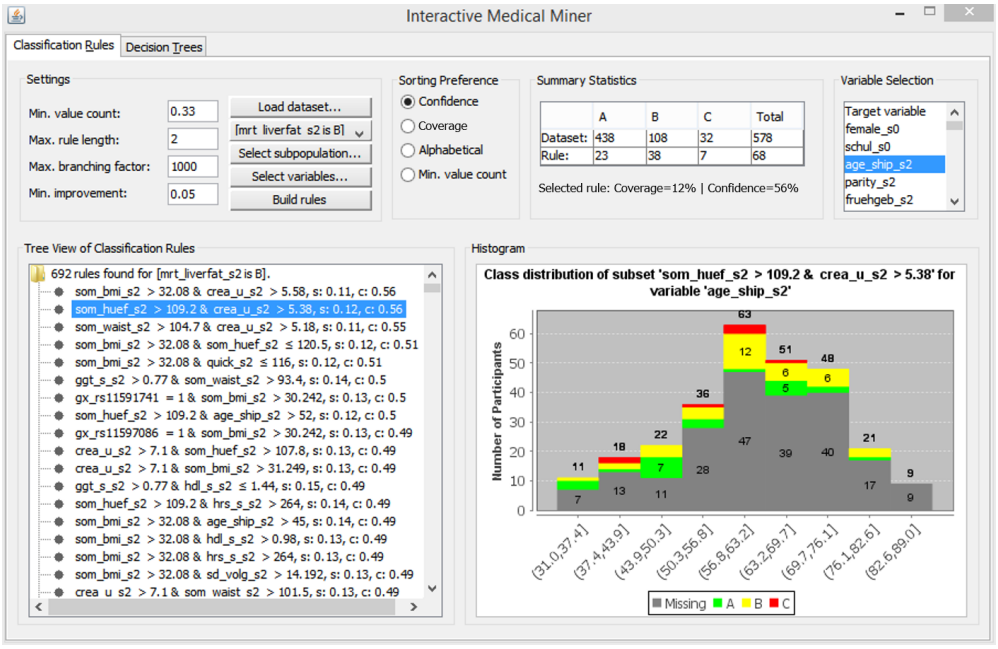

Figure 3.6: The Interactive Medical Miner: classification rules are discovered for class B and shown in the bottom left panel. For the selected rule som_huef_s2 $>$ 109 & crea_u_s2 $>$ 5.38 $\longrightarrow$ mrt_liverfat_s2 = B, the distribution of the participants covered by the rule among all three classes is shown in absolute values (top middle panel) and as histogram (bottom right panel) with respect to age (top right panel).

threshold. The position of a condition within the antecedent indicates at which refinement step it was added to the rule candidate. For example, the first condition som_huef_s2 > 109 with a confidence of 44/107 = 0.41 was expanded by the second condition crea_u_s2 > 5.38, because the gain in confidence exceeds the minimum improvement threshold, i.e., 38/68 - 44/107 = 0.15 > 0.05. This rule, however, cannot be further expanded because the maximum rule length is set to 2. The maximum branching factor was set conservatively to 1000, to prevent potentially interesting rules from not being generated due to small beam width. The parameter can be lowered interactively, if the number of discovered rules is to high or rule induction takes too long.

The output list of an execution run (area below the "Settings") is scrollable and interactive. When the expert clicks on a rule, the top middle area "Summary Statistics" is updated. The first row shows the distribution of the cohort participants among the classes for the whole dataset, while the second row shows how the participants covered by the rule (column "Total" in the second row) are distributed among the classes. Hence, the expert can specify the discovery of classification rules for one of the classes and then study how often the antecedent of each rule appears among the participants in the other classes. A rule that covers most of the participants of the selected class (class B in Figure 3.6) is not necessarily interesting, for example, if it covers also a high number of participants of the other classes. The rule som_huef_s2 > 109 & crea_u_s2 > 5.38 $\longrightarrow$ mrt_liverfat_s2 = B covers 68 participants in total, 38 of them having class B. To lower the number of covered participants from other classes , i.e., to increase confidence, the user can lower the minimum value count, to allow for the generation of rules with lower sensitivity, but more homogeneity with respect to the selected class.

Some of the data might be incomplete. For example, not all participants in the cohort have been subjected to liver MRI. Hence, it is also of interest to know the distribution of the unlabeled participants who support the antecedent of a given rule. The "Histogram" panel can be used to this purpose: the expert chooses a further feature from the interactive area "Variable selection" in the upper right panel and can then see how the values of these variable are distributed among the study participants - both the labeled ones and the unlabeled ones; the latter are marked as "Missing" in the color legend. For plotting the histograms we make use of the free Java chart library JFreeChart [58]. Numeric variables are discretized using "Scott's rule" [162] as follows: let $X_{s(r)}$ be the set of values for a numeric variable X with respect to the cover set $s(r)$. The bin width h is then calculated as $h(X_{s(r)}) = \frac{\max X_{s(r)} - \min X_{s(r)}}{3.49 \mathrm{sd}_{s(r)}} \cdot |s(r)|^{\frac{1}{3}}$.

If the expert does not choose any variable, the target variable is used by default, and only the distribution of the labeled participants is visible. The histogram in Figure 3.6 shows the age distribution of both labeled and unlabeled participants covered by our example rule som_huef_s2 > 109 & crea_u_s2 > 5.38 $\longrightarrow$ mrt_liverfat_s2 = B. The value distribution among the labeled participants reveals that ageing might be a risk factor for the displayed subpopulation as the likelihood of class B increases with higher age. This visual finding suggests adding the condition age_ship_s2 > 56.8 to the rule's antecedent. Indeed, the confidence of this more specific rule increases from 38/68 = 0.56 to 27/40 = 0.675. However, since the sensitivity reduces from 38/108 = 0.352 to 27/108 = 0.250, the minimum value count threshold is no longer satisfied. Hence, the visualization of the participants' statistics for selected rules can deliver indications on subpopulations that should be monitored closer, and hints on how to alter the algorithm parameters for subsequent runs, in our example, to lower the minimum value count to 0.25 and to increase the maximum rule length to 3.

Table 3.2: Best decision trees for the three partitions: best separation is achieved in F : age > 52; PartitionM is the most heterogeneous one, the performance values are lowest.

Partition	Variant	Sensitivity (%)	Specificity (%)	F1 score (%)
F : age > 52	Oversampling	63.5	93.9	81.5
PartitionF	InfoGain	52.4	94.9	69.7
PartitionM	CostMatrix	38.3	86.3	53.0

3.3 Experiments and Findings

3.3.1 Results of Decision Tree Classifiers

For the evaluation of decision tree classifiers, we consider accuracy, i.e., the ratio of correctly classified participants, sensitivity and specificity, and the F_1 score, i.e., the harmonic mean between precision and recall. For **specificity**, precision and recall, we consider the two classes B and C together as positive class.

Oversampling achieved best performance with an accuracy of ca. 80% and an F_1score of 62%. The best decision trees were found for partition F : age > 52, followed by those for PartitionF, then PartitionM. **The large discrepancy between accuracy and F_1score appears also in the models of the partitions, underlying that accuracy scores are unreliable in such a skewed distribution. Therefore, we do not report on accuracy hereafter.**

On partition F : age > 52, the overall best decision tree is achieved by the oversampling variant. On the larger PartitionF, best performance was achieved by the decision tree produced with the InfoGain variant, while the best decision tree on PartitionM was built with the CostMatrix variant. Sensitivity and specificity values for these trees are shown in Table 3.2, while the trees themselves are depicted in Figures 3.7 - 3.9 respectively and discussed in Section 3.3.3.

Table 3.2 indicates that the decision tree variants perform differently on different partitions. Oversampling is beneficial for F : age > 52, because it partially compensates the skew problem. As PartitionM has the most heterogeneous class distribution out of all paritions, all variants perform relatively poor on it. Hence, we expected most insights from the decision trees on F : age > 52 and PartitionF, where better separation is achieved.

3.3.2 Discovered Classification Rules

While the classification rules found by HotSpot on the whole dataset were conclusive for class A but not for the positive classes B and C, we omit reporting these rules as they are not useful for diagnostic purposes. The classification rules found on the partitions were more informative. However, classification rules with only one feature in the antecedent had low confidence. To ensure high confidence, we restricted the output on rules with at least two features in the antecedent. To ensure a still high coverage, we allowed for at most three features. A selection of high confidence and high coverage rules for each partition and class are shown in Tables 3.3 – 3.5, respectively. We describe the most important features in the antecedent of these rules in the next subsection, together with the most important features of the best decision trees.

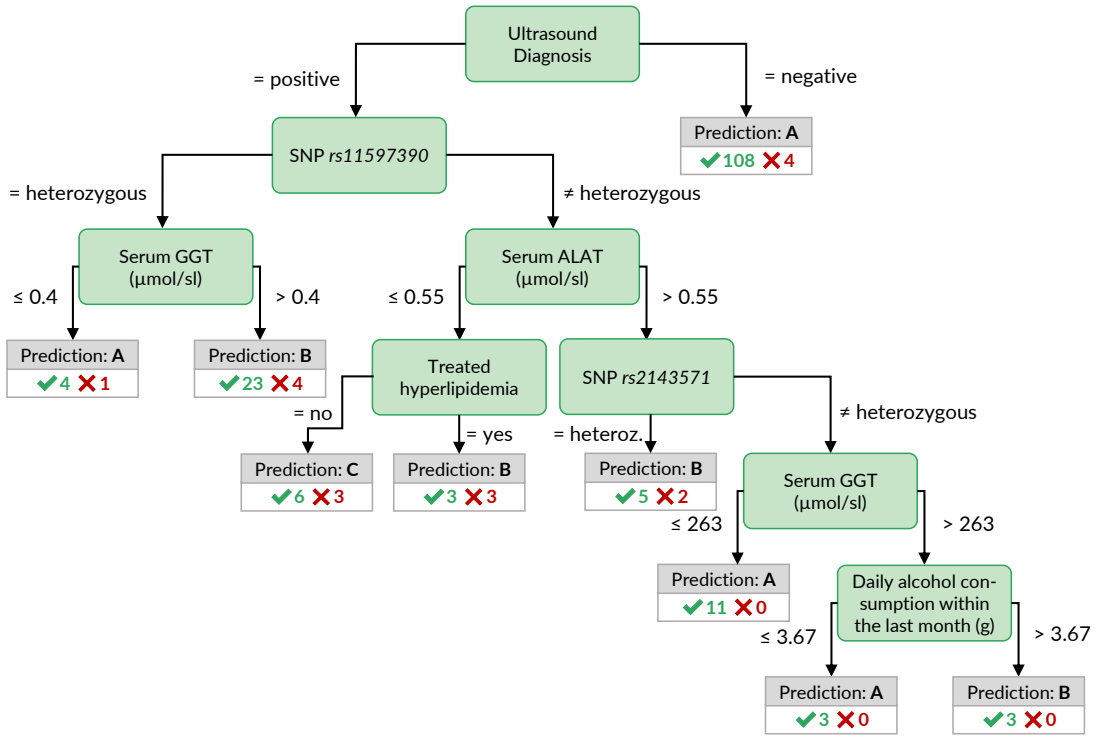

Figure 3.7: **Best decision tree for** F : age $>$ 52, achieved by the variant *Oversampling*.

Table 3.3: Best HotSpot classification rules (*maxLength* = 3) for PartitionF (excerpt).

Rule antecedent			Cov	Sup		Conf
Variable 1	**Variable 2**	**Variable 3**	**Abs**	**Abs**	**Rel [%]**	**Rel [%]**
Target class: A						
som_waist_s2 $\leq$ 80	–	–	132	132	52	100
som_bmi_s2 $\leq$ 24.82	–	–	109	109	43	100
som_huef_s2 $\leq$ 97.8	–	–	118	117	46	99
stea_s2 = 0	–	–	218	214	84	98
stea_alt75_s2 = 0	–	–	202	198	78	98
Target class: B						
stea_s2 = 1	gx_rs11597390 = 1	age_ship_s2 $>$ 59	20	17	40	85
stea_alt75_s2 = 1	hrs_s_s2 $>$ 263	age_ship_s2 $>$ 59	20	17	40	85
stea_alt75_s2 = 1	hrs_s_s2 $>$ 263	ldl_s_s2 $>$ 3.22	20	17	40	85
stea_s2 = 1	age_ship_s2 $>$ 66	tg_s_s2 $>$ 1.58	17	14	33	82
stea_s2 = 1	age_ship_s2 $>$ 64	hrs_s_s2 $>$ 263	17	14	33	82
Target class: C						
gluc_s_s2 $>$ 7	tsh_s2 $>$ 0.996	–	6	6	35	100
som_bmi_s2 $>$ 38.42	age_ship_s2 $\leq$ 66	asat_s_s2 $>$ 0.22	6	6	35	100
som_bmi_s2 $>$ 38.42	sleeph_s2 $>$ 6	blt_beg_s2 $\leq$ 38340	6	6	35	100
som_bmi_s2 $>$ 38.42	sleeph_s2 $>$ 6	stea_s2 = 1	6	6	35	100
hrs_s_s2 $>$ 371	sleepp_s2 = 0	ggt_s_s2 $>$ 0.55	6	6	35	100

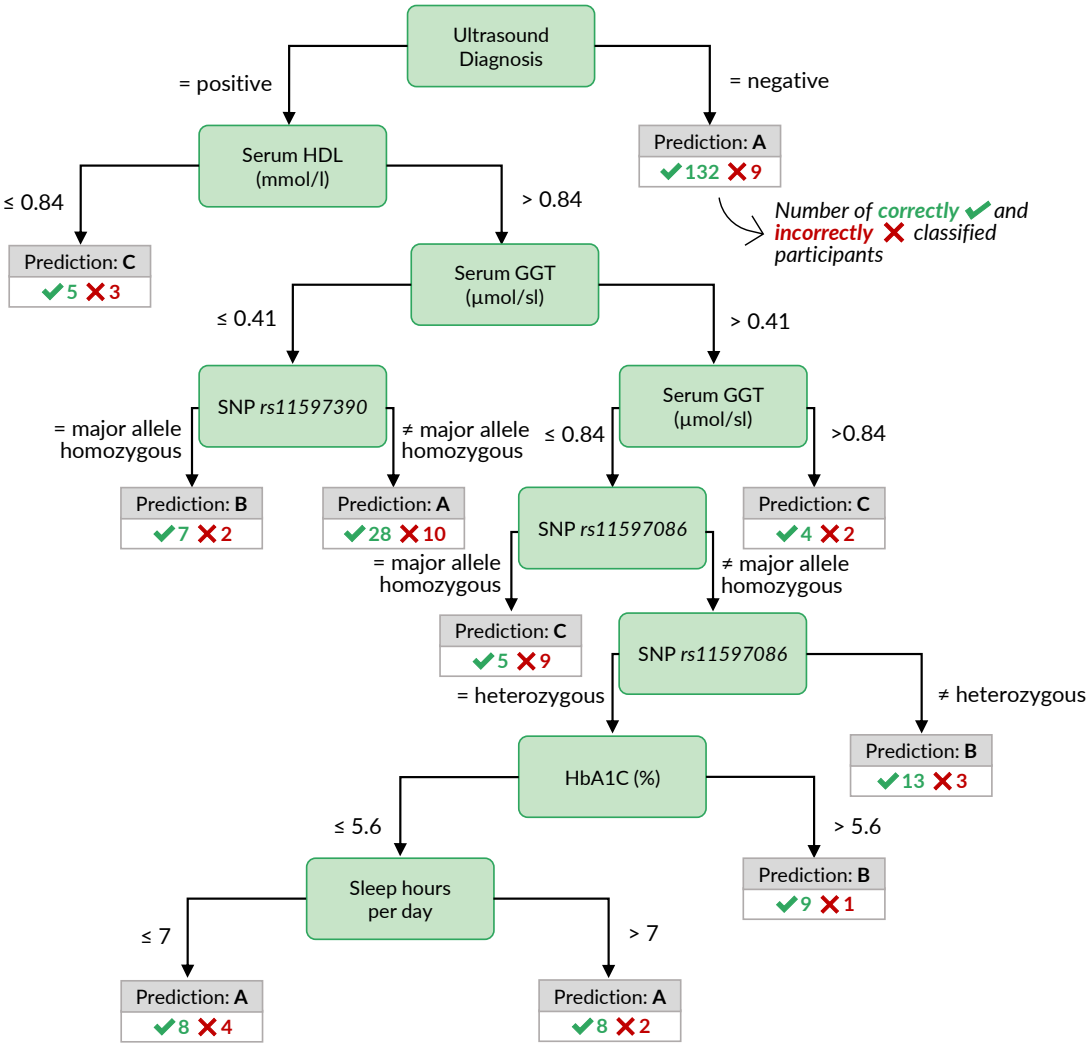

Figure 3.8: **Best decision tree for** PartitionF, achieved by the variant *InfoGain*.

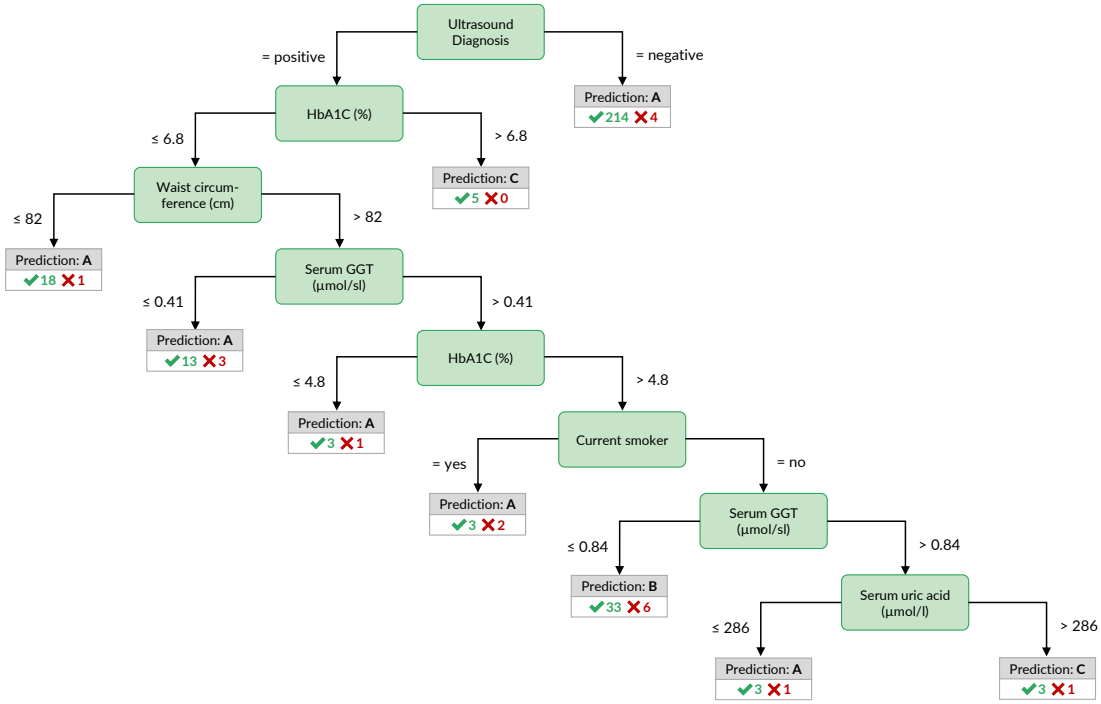

Figure 3.9: **Best decision tree for** PartitionF, achieved by the variant *InfoGain*.

Table 3.4: Best HotSpot classification rules ($maxLength = 3$) for $\mathsf{F} : \mathsf{age} > 52$ (excerpt).

Rule antecedent			Cov	Sup		Conf
Variable 1	**Variable 2**	**Variable 3**	**Abs**	**Abs**	**Rel [%]**	**Rel [%]**
Target class: A						
crea_u_s2 $\leq$ 5.39	stea_s2 = 0	–	75	75	57	100
crea_u_s2 $\leq$ 5.39	stea_alt75_s2 = 0	–	72	72	55	100
som_waist_s2 $\leq$ 80	–	–	54	54	41	100
som_bmi_s2 $\leq$ 24.82	–	–	50	50	38	100
crea_u_s2 $\leq$ 5.39	ggt_s_s2 $\leq$ 0.43	–	50	50	38	100
Target class: B						
stea_s2 = 1	ggt_s_s2 > 0.48	ggt_s_s2 $\leq$ 0.63	15	15	38	100
stea_s2 = 1	gx_rs11597390 = 1	hdl_s_s2 $\leq$ 1.53	20	19	48	95
stea_s2 = 1	gx_rs11597390 = 1	fib_cl_s2 > 3.4	15	14	35	93
crea_s_s2 $\leq$ 61	som_waist_s2 > 86	stea_s2 = 1	15	14	35	93
stea_s2 = 1	gx_rs11597390 = 1	hrs_s_s2 > 261	20	18	45	90
Target class: C						
som_bmi_s2 > 38.42	age_ship_s2 $\leq$ 66	–	4	4	33	100
som_bmi_s2 > 38.42	stea_alt75_s2 = 3	–	4	4	33	100
som_huef_s2 > 124	stea_alt75_s2 = 3	–	4	4	33	100
som_waist_s2 > 108	gluc_s_s2 > 6.2	–	4	4	33	100
stea_alt75_s2 = 3	som_bmi_s2 > 37.32	–	4	4	33	100

Table 3.5: Best HotSpot classification rules ($maxLength = 3$) for PartitionM (excerpt).

Rule antecedent			Cov	Sup		Conf
Variable 1	**Variable 2**	**Variable 3**	**Abs**	**Abs**	**Rel [%]**	**Rel [%]**
Target class: A						
stea_alt75_s2 = 0	–	–	106	101	55	95
stea_s2 = 0	–	–	138	131	72	95
ggt_s_s2 ≤ 0.52	–	–	79	73	40	92
hrs_s_s2 ≤ 310	ggt_s_s2 ≤ 0.77	–	81	74	40	91
som_waist_s2 ≤ 90.8	–	–	79	72	39	91
Target class: B						
som_huef_s2 > 108.1	age_ship_s2 > 39	crea_u_s2 > 7.59	28	22	33	79
som_bmi_s2 > 32.29	hdl_s_s2 > 0.94	ATC_C09AA02_s2 = 0	29	22	33	76
som_bmi_s2 > 32.29	hgb_s2 > 8.1	gout_s2 = 0	29	22	33	76
som_waist_s2 > 109	sleeph_s2 ≤ 8	jodid_u_s2 > 9.44	29	22	33	76
som_huef_s2 > 108.1	hdl_s_s2 > 0.97	crea_u_s2 > 5.38	29	22	33	76
Target class: C						
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	quick_s2 > 59	6	6	40	100
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	chol_s_s2 > 4.3	6	6	40	100
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	fib_cl_s2 > 1.9	6	6	40	100
ggt_s_s2 > 1.9	crea_s_s2 ≤ 90	crea_u_s2 > 4.74	6	6	40	100
ggt_s_s2 > 1.9	tg_s_s2 > 2.01	som_waist_s2 > 93.5	6	6	40	100

3.3.3 Important Features for Each Subpopulation

The most important features in the decision trees of Figures 3.7 - 3.9 are those closer to the root. For better readability, the tree nodes in the figures contain short descriptions instead of the original variable names. In all three decision trees, the root node is the ultrasound diagnosis variable stea_s2. A negative ultrasound diagnosis points to the negative class A, but a positive ultrasound diagnosis does not directly lead to one of the positive classes B and C. The decision trees of the three partitions differ in the nodes placed near the root.

In the best decision tree of PartitionF (cf. Figure 3.8) we observe that if the ultrasound report is positive *and* the HbA1C concentration is more than 6.8%, the class is C. The classification rules with high coverage and confidence in Table 3.3) point to further interesting features: a waist circumference of at most 80 cm, a BMI of no more than 24.82 kg/m^2, a hip circumference of 97.8 cm or less characterize participants of the negative class. All 6 participants with a serum glucose concentration greater than 7 mmol/l *and* a TSH concentration greater than 0.996 mu/l belong to class C. Further, severe obesity (a BMI value of more than 38.42 kg/m^2 points to class C with high confidence – but only in combination with other variables.

In contrast to the best tree for PartitionF, the best decision tree for the subpartition F : age > 52 (cf. Figure ~3.7) also contains nodes with SNPs, indicating potentially genetic associations to fatty liver for these participants. Classification rules with high coverage and confidence for class B also contain SNPs, as can be seen in Table ~3.4. Similarly to PartitionF, high BMI values point to a positive class when combined with other features: in Table 3.4, we see that all four participants with stea_alt75_s2 = 3 (i.e. a positive ultrasound diagnosis combined with a critical ALAT value) and a BMI larger than 38.42 kg/m^2 belong to class C. A similar association holds for stea_alt75_s2 = 3 combined with a high waist circumference (> 124 cm). 19 out of 20 participants in class B having a positive ultrasound diagnosis, a genetic marker gx_rs11597390 = 1 and a serum HDL concentration of at most 1.53 mmol/l.

The role of the ultrasound report in predicting the negative class is the same for PartitionM (cf. Figure 3.9 as for PartitionF. As with the best tree for F : age $>$ 52, the best tree for PartitionM contains nodes with SNPs and serum GGT value ranges. Such features are also in the antecedent of top Hotspot rules (cf. Table ~3.5): a Serum GGT concentration of more than 1.9,μmol/sl in combination with creatinine concentration of at most 90 mmol/l or a thromboplastin time ratio (quick_s2) of more than 59% point to class C. Similarly, positive ultrasound diagnosis and a serum HDL concentration not exceeding 0.84 mmol/l point to class C.

The decision trees and classification rules give insights into features that seem diagnostically important. However, the medical expert needs additional information to decide whether a feature is worth further investigation. In particular, decision trees highlight the importance of a feature only in the context of the subtree it is located; a subtree describes a subpopulation that is usually very small. In contrast, classification rules return information on larger subpopulations. However, these subpopulations may overlap; for example, the first four rules on class C for PartitionM (cf. Table 3.5) may refer to the same 6 participants. Moreover, unless a classification rule has a confidence close to 100 %, there may be participants in the other classes that also support it. Hence, to decide whether the features in the rule's antecedent deserve further investigation, the expert also needs insights on the rule's statistics for the other classes as well. To assist the expert in this task, we propose the Interactive Medical Miner, a tool that discovers classification rules for each class *and* delivers information on the statistics of these rules for all classes.

3.3.3.1 Enhancements

In [136], we added functionalities to specify a (sub-)cohort of the dataset, for example, to focus on the subset of female participants. In an additional window, the expert can specify filtering queries in the form of `Variable Operator Value`. The defined restrictions are shown in a table where the user can undo previous constraints. Further, the user can (de-)select variables for model generation. For example, the expert might exclude a variable that is known to be highly correlated with another variable that is already considered for model learning.

[159] added panels that contain tables displaying additional rule statistics such as lift and p-value. Further, a mosaic chart juxtaposes the class distributions of a subpopulation and its "complements" which are subgroups of participants who do not match one or both of the conditions of the length-2 rule which describes that subpopulation.

In this paper, we study SD on a binary target variable problem (positive or negative outcome). A subpopulation of r, $s(r)$ is the set of instances that satisfy the antecedent of a rule r, also known as cover set of r. % % %A subpopulation comprises of instances that fulfill the logical condition of a rule's antecedent; ae a rule is described as:
% %Further, $|s(r)_{T=v}|$ is the number of instances that additionally belong to the value for the target variable in $s(r)$, the support set; n is the total number of instances in the dataset; $n_{T=v}$ is the total number of instances which exhibit the target variable value of interest. % %

The conjunction of feature - feature value pairs of left to the arrow constitute the rule's *antecedent* (or left hand site), while the pair of target feature - target feature value right to the arrow is the rule's *consequent* (or right hand side). The rule in Equation (3.2) is a *classification rule* referring to class B.

Features with continuous values (e.g. age, waist) are restricted within a specific range. This range is chosen by the algorithm in such a way as to maximize the support of the rule within the cohort and the confidence of the rule's consequent given the antecedent. The first rule in the example belongs to the results on PartitionF (see Table 3.3, described in the next section): out of the 20 participants supporting the antecedent, 17 belong to class B. Note that we use the

expressions *"[number] participants support rule [ruledescription] "* and *"[number] participants support the rule's* antecedent*"*.

For classification rule discovery we use the Weka algorithm HotSpot. For each class, this algorithm determines the rules with the best confidence *and* the optimal boundary values for the features in the antecedent. We have wrapped the algorithm into a mechanism that selects for each class only rules supported by at least τ participants. In our experiments, we use $\tau = \frac{1}{3}$, but our interactive tool (section **??**) allows the expert to set this threshold freely.

3.4 Conclusions on ...

[144]

Chapter 4

Identifying Distinct Subpopulations

Brief Chapter Summary

We study redundancy in large rule sets describing subpopulations. We present a workflow that extracts a smaller number of representative rules. These rules are selected to avoid instance overlap as much as possible, thus covering different concepts in the data space. We evaluate our workflow on samples of SHIP with hepatic steatosis and goiter as target variables.

This chapter is partly based on:
Uli Niemann, Myra Spiliopoulou, Bernhard Preim, Till Ittermann, and Henry Völzke. "Combining Subgroup Discovery and Clustering to Identify Diverse Subpopulations in Cohort Study Data". In: *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2017, pp. 582-587. DOI: 10.1109/CBMS.2017.15.

Epidemiologists search for significant relationships between risk factors and outcome in large and heterogeneous datasets that encompass participant health information gathered from questionnaires and medical examinations. In the previous chapter, we described a expert-driven workflow that can help epidemiologists to automatically detect such relationships in form of classification rules, which are descriptions of risk factors and predictive value ranges in for a specific subpopulation and an outcome of interest. However, rule induction algorithms often produce large and overlapping rule sets requiring the expert to manually pick the most interesting rules and to remove less interesting or redundant rules. This post-filtering step is time-consuming and tedious. This chapter presents a clustering-based algorithm that hierarchically reorganizes large rule sets and summarizes all important concepts while maintaining *distinctiveness* between the clusters. For each cluster, a representative rule is shown to the expert who in turn can drill-down to other cluster members. We evaluate our algorithm on two subsets of SHIP where the outcomes hepatic steatosis and goiter serve as target variable, respectively. Further, we report on effectiveness of our algorithm and we present selected subpopulations.

This chapter presents SD-Clu, an approach that combines subgroup discovery with clustering to return a set of k representative classification rules. Building up on a set of potentially highly

overlapping rules generated by a SD algorithm, we leverage hierarchical agglomerative clustering to find groups of rules that cover different sets of instances. For each cluster, we nominate one rule as the group's representative that exhibits best trade-off between rule confidence and coverage towards the target variable. We define similarity between a pair of subgroups based on the fraction of mutually covered instances and individually covered instances. Rules covering (almost) the same instances are likely to be condensed into the same cluster and thus more likely to be represented by a proxy rule. We evaluate our algorithm on two samples from *SHIP* investigating the diseases hepatic steatosis and goiter.

Section 4.1 serves as motivation for the related field of subgroup discovery and redundancy of classification rules. Section 4.2 presents our SD-Clu algorithm that generates distinct rules. Section 4.3 contains the experimental setup. In Section 4.4, we describe our evaluation results. In Section 4.4.1, we discuss our findings regarding hepatic steatosis and goiter on the SHIP data. In Section 4.5, we summarize our contributions.

4.1 Motivation and Comparison to Related Work

Subgroup discovery (SD) algorithms aim to uncover *interesting* relationships between one or more conditions (variables and value ranges) and a target variable in the form of classification rules [79, 6]. Compared to more accurate but predominantly opaque black-box models such as neural networks, support vector machines, or random forests, SD algorithms provide higher confidence and interpretability of results, making them highly suitable for domain expert-guided subpopulation discovery. SD algorithms have been used in several medical studies where descriptive knowledge needed to be inferred, e.g., to extract potential drug targets from multi-relational data sources for the treatment of dementia [134], to identify predictive auditory-perceptual, speech-acoustic, and articulatory-kinematic features of preschool children with speech sound disorders [183], and to discover discriminative features in patient subpopulations with different admission times to psychiatric emergency departments [24].

However, SD methods often yield large sets of rules that domain experts are not willing to tediously go through all of them to manually separate interesting from irrelevant and redundant rules. A common observation is that there are groups of rules that cover almost the same set of instances, as shown in Figure 4.1. Instead of presenting *all* rules found by an SD algorithm at once, we propose to organize rule sets hierarchically so that the domain expert is able to explore a compact set of different concepts, equipped with mechanisms to drill down to specific rules of interest.

Similar to the HotSpot algorithm described in Section 3.2.4, popular SD algorithms such as SubgroupMiner [104], SD [55] and CN2-SD [119] use a fixed beam width to limit the number of further expanded subgroup candidates at each iteration. A post-pruning step can be applied to reduce the cardinality of the rule set – e.g., to return the *top-k* rules – using a quality criterion such as the weighted relative accuracy or the p-value of a statistical test. Even when both beam-width search and top-k pruning are applied, the result often still contains redundant rules. This is due to the correlation between the (non-target) variables, which leads to a large number of variations of a given finding, cf. Figure 4.2 for an illustrative example. In particular, top-k pruning leads to different variations of the same concepts (i.e. large instance overlap) [120].

Unlike SD, which generates multiple rules that overlap in terms of their coverage sets, predictive rule learning algorithms such as CN2 [31] or RIPPER [32] are designed to generate rules that capture different spaces in the data. They work iteratively according to a divide-and-conquer strategy [54]: First, a rule that maximizes the quality function of the algorithm is generated on the set of instances not yet covered. Second, all covered instances are removed from the training

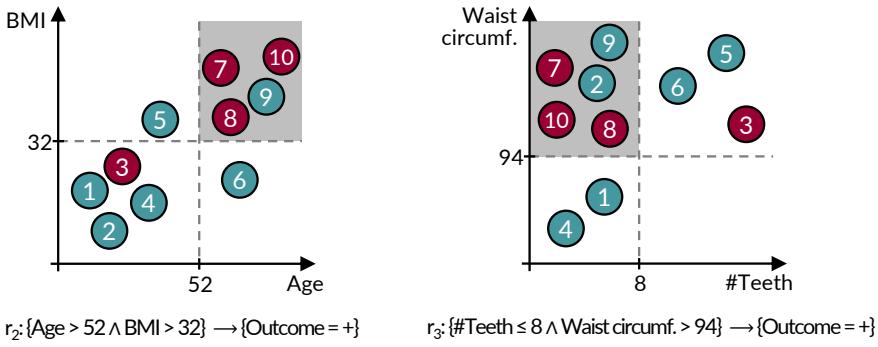

Figure 4.1: **Example of two *redundant* rules.** Both r_1 and r_2 cover the instances 7,8,9,10; r_2 additionally covers instance 2. The cover set overlap is due to the high correlation between *#Teeth* and *Age*, and between *BMI* and *Waist circumference*. Both rules describe the same concept, i.e., *elderly overweight people have a higher risk of the outcome*.

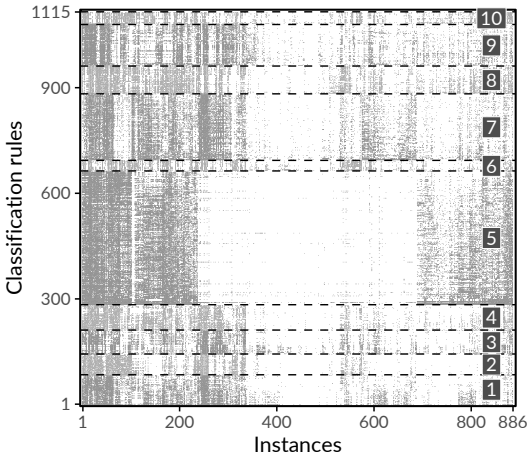

Figure 4.2: Graphical representation of 1115 HotSpot rules found on SHIP data on 886 labeled instances. A gray cell indicates that the instance at that x-position is covered by the rule at the associated y-position. Instances and rules are sorted by agglomerative hierarchical clustering. When partitioning into 10 clusters based on the covered instances, the rules in each cluster describe similar subpopulations.

set. This process of rule induction and removal of instances from the training set is repeated until all instances are covered by at least one rule. The output of this algorithm is often a decision list. To classify an instance, the prediction of the first rule that covers the instance is used. While these algorithms avoid the problem of rule redundancy, important concepts may remain uncovered. For example, the algorithm might induce a rule that includes instances with a high BMI, but it might not be able to find a slightly weaker association with income because those instances are immediately removed from the instance candidate space. In addition, the coverage of rules generally decreases with each iteration. Rules with low coverage are negligible in epidemiological studies because they may represent spurious correlations of the study sample.

Instead of simply eliminating covered instances from rule induction of subsequent iterations, *weighted coverage approaches* take into account how many times instances have been covered so far in the rule candidate expansion step [119]. While this leniency over removing instances allows for a larger number of rules, a new hyperparameter is introduced to control the tradeoff between reusing already covered instances and minimum rule confidence. This parameter is unintuitive and difficult to set, especially for domain experts. Moreover, both traditional predictive rule learning algorithms and their weighted covering extensions introduce order dependencies between rules: a rule depends on all previous rules in the rule list and the instances of the target variable it covers, and it may not make sense to interpret a single rule.

4.2 Finding Distinct Classification Rules

In this section, we present our algorithm SD-Clu, which combines subgroup discovery with clustering to return a set of k distinct classification rules. The algorithm consists of three main steps. First, the SD algorithm generates a set of potentially highly overlapping rules. Using hierarchical agglomerative clustering, this set of rules is then grouped into a set of distinct rule clusters covering different sets of instances. For each cluster, the rule that has the best tradeoff between confidence and coverage of the target variable is appointed as the representative of the group. Rules covering the same instances are grouped in the same cluster and are therefore represented by the same proxy rule.

4.2.1 Rule clustering

We use the same notation for classification rule discovery as in Section 3.2.3. Agglomerative hierarchical clustering iteratively merges similar instances to clusters, in a bottom-up way. The order of merging two clusters depends on the linkage strategy: in *complete linkage* the distance between two clusters is defined as the maximum distance between any two of their instances. The pair of clusters minimizing this maximum distance are selected for merging. A dendrogram is a tree representation of this stepwise process. Optionally, a parameter k can be specified to obtain a specific partitioning. We define rule similarity for clustering on the basis of the mutually covered instances as an adaption of the Sørensen–DICE coefficient [38]. The distance between two rules r_1, r_2 with corresponding subpopulations $s(r_1)$, $s(r_2)$ is given as

$$\mathrm{dist}(r_1, r_2) = 1 - \frac{2 \cdot |s(r_1) \cap s(r_2)|}{|s(r_1)| + |s(r_2)|}. \tag{4.1}$$

The number of clusters can be specified as parameter k. Alternatively, we describe an approach that derives an appropriate k from the rule set using a cluster quality function such as the Silhouette coefficient. For a clustering ξ and a set of rules R, the Silhouette coefficient is calculated as

$$\mathrm{Silh}(R, \xi) = \frac{1}{|R|} \sum_{r \in R} \frac{b(r) - a(r)}{\max\{a(r), b(r)\}} \tag{4.2}$$

where

$$a(r) = \frac{\sum_{y \in Y} \text{dist}(r, y)}{|Y| - 1} \tag{4.3}$$

is the average dissimilarity between r and the other rules in the cluster $Y \in \xi$ which contains r, while

$$b(r) = \frac{\sum_{y \in Z} \text{dist}(r, y)}{|Z|} \tag{4.4}$$

is the average dissimilarity between r and the rules in the cluster $Z \in \xi$ which is the closest to the cluster Y containing r. Then, we traverse the dendrogram bottom-up, compute the Silhouette for each set of clusters ξ and select as ξ_{opt} the set of clusters with the best Silhouette value. The optimal number of clusters is then the cardinality $|\xi_{opt}|$. Finally, we map each cluster $Y \in \xi_{opt}$ to a representative rule. To do so, we invoke the rule interestingness measure *Weighted Relative Accuracy* (WRA hereafter) which balances coverage and confidence gain and is defined as

$$WRA(r) = Cov(r) \cdot \left(Conf(r) - \frac{n_{T=v}}{N} \right) \tag{4.5}$$

where N is the total number of instances in the dataset and $n_{T=v}$ is the number of instances with the target variable value of interest. We compute WRA for each rule $r \in Y$ and select as cluster proxy $cp(Y)$ the rule for which WRA is maximum.

4.2.2 Representativeness of a set of cluster proxies

The rule proxies should be a good representation of the total rule set. Thus, each instance should be covered equally often by the cluster proxies compared to the total rule set. Hence, we define representativeness as difference between the average fraction of proxy rules the instances are covered by and the total average fraction of rules the instances are covered by. If the difference between the two ratios is small, then the representativeness of the cluster proxy rules is high.

Typically, the set of rule clusters ζ for a set of rules R will be the optimal set of clusters, as described in the previous subsection, but it can be any set of clusters chosen by the user, as long as it contains all rules in R. For ζ, let $R_\zeta = \{cp(Y) | Y \in \zeta\}$ denote the set of cluster proxy rules. To quantify how representative such a set of rules is, we proceed as follows. First, let $U \subseteq R$ be an arbitrary subset of the complete set of rules, and let x be an instance. The *coverage rate* of x towards U is calculated as

$$covRate(x, U) = \frac{\sum_{r \in U} isCovered(x, r)}{|U|} \tag{4.6}$$

where $isCovered(x, r)$ is equal to 1, if r covers x, i.e., $x \in s(r)$, and 0 otherwise. We observe that for a set of rules U, an instance x cannot be covered by more than $|U|$ rules. Let R_x be the set of rules that cover instance x, i.e., for $R_x = \{r \in U | isCovered(x, r) = 1\}$. For the whole set of instances X we create bins:

$$bin_i(U) = \{x \in X | |R_x| = i\}. \tag{4.7}$$

Further, let $bin_0(U) = \{x \in X | \forall r \in U : isCovered(x, r) = 0\}$. Evidently, an instance x can be covered by $0, 1, \ldots, |U|$ rules, i.e. $covRate(x, U)$ can take one of $|U| + 1$ values. In contrast, $covRate(x, R)$ can take one of $|R| + 1$ values, a usually much larger number. We therefore map the possible values of $covRate(x, R)$ into the much smaller set of possible values by rounding, computing:

$$adjCovRate(x, U, R) = \frac{\lfloor covRate(x, R) \cdot |U| \rceil}{|U|} \tag{4.8}$$

where $\lfloor \rceil$ is the rounding operator. Then, for the complete set of instances X, a set of induced rules R, the clustering ζ over R and the set of cluster proxy rules R_ζ, the *representativeness* of R_ζ is defined as

$$representativeness(R_\zeta, R) = 1 - \frac{1}{|X|} \sum_{x \in X} |adjCovRate(x, U, R) - covRate(x, R_\zeta)|. \quad (4.9)$$

4.3 Experimental Setup

Datasets. To evaluate our method, we used data from the SHIP study. For a description of SHIP, see Section 2.6.1. We considered hepatic steatosis (see section 3.2.1) and goiter as target variables. For the sample `HepStea`, we derived a binary outcome variable by discretizing the liver fat concentration obtained from the MRI report so that study participants with a concentration no greater than 10% were assigned to the negative class and values greater than 10% were assigned to the positive class indicating the presence of the disease. Of 886 participants for whom the MRI report from SHIP-2 was available at that time, 694 (78.3%) were negative and 192 (21.7%) were positive. We considered 99 variables selected exclusively from SHIP-0 to assess their long-term effects as expressed 10 years later in SHIP-2. For the `Goiter` sample, the outcome variable was derived by thyroid ultrasound. The presence of goiter was defined for a thyroid volume greater than 18ml in women and 25ml in men [67]. Of the 4400 participants for whom the outcome variable is available in TREND-0, 3010 belong to the negative class (68.4%) and 1390 (31.6%) to the positive class. Apart from the target variable, we use a total of 182 variables that were pre-selected by a medical expert as potential risk factors.

SD algorithms. For subgroup discovery, the two algorithms HotSpot [70] (cf. section 3.2.4 and SD-Map [8] are used. SD-Map is an exhaustive algorithm that adapts the popular FP-Growth association rule learning method [73]. Rules that fall below a minimum coverage threshold are pruned. A deep-first search is performed for candidate generation. Rules are ranked according to a user-defined quality function. We used the implementation from the VIKAMINE framework [7]. The implementation of SD-Map only supports categorical variables. Therefore, each numeric variable was discretized using the minimum description length based approach of Fayyad and Irani [46]. For SD-Map, we set the minimum coverage threshold to 0.05 to avoid overfitting rules that are too small. We use WRA as a quality function and define a minimum threshold of 0.025. For HotSpot, we set the support threshold of a rule to be above 0.05. The beam width is set to 500. In addition, to avoid rather meaningless literals, we restrict the extension of a rule body with another literal to a relative confidence gain of at least 0.3. To avoid having many overly specific rules that cover only a small number of study participants, we limit the length of a rule body, i.e., the number of literals to 3.

4.4 Results

In Figure 4.3, we show the optimal number of clusters for each combination of study sample and SD algorithm. Table 4.1 shows the optimal k and the ratio of proxy rules vs. total number of rules. For example, clustering with optimal silhouette coefficient for the algorithm HotSpot on `Goiter` has 76 clusters and thus 76 rule cluster proxies (cf. Table 4.1), which is only 21.3% of the total number of rules. Thus, if only the cluster proxies are displayed to the expert at the beginning, the time needed to check the rules is reduced.

The optimal cardinality of $|\zeta_{opt}$ shown in Figure 4.3 could be used as a suggestion, but the expert is free to specify the number of rules they wishes to obtain. For example, if the expert

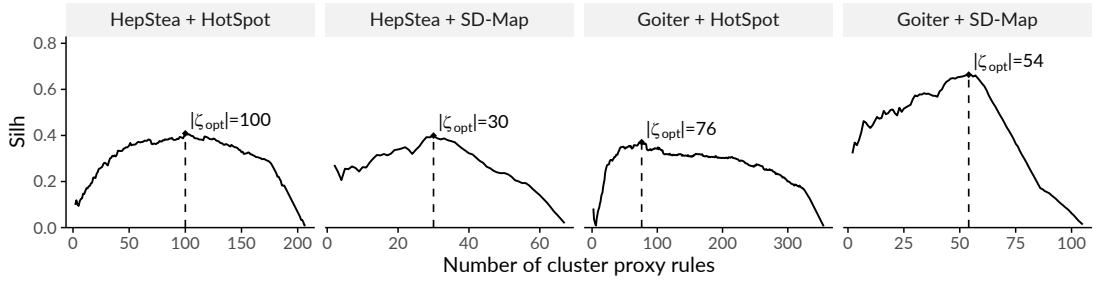

Figure 4.3: **Silhouette coefficients ($Silh$) of SD-Clu using complete linkage for each combination of dataset and algorithm.** The cardinality of the clustering with the highest $Silh$ score ($|\zeta_{opt}|$) is indicated by a dashed vertical line.

Table 4.1: **Statistics of best runs per dataset and algorithm.** Number of rules $|R|$, optimal Silhouette coefficient $Silh(\zeta_{opt})$, the corresponding cardinality of the optimal clustering $|\zeta_{opt}|$ and percentage of cluster proxies relative to the total number of rules for every combination of data sample and SD algorithm.

Dataset	Algorithm	$\lvert R\rvert$	$Silh(\zeta_{opt})$	$\lvert\zeta_{opt}\rvert$	$\frac{\lvert\zeta_{opt}\rvert}{\lvert R\rvert}$[%]
HepStea	HotSpot	208	0.41	100	48.1
HepStea	SD-Map	68	0.40	30	44.1
Goiter	HotSpot	356	0.37	76	21.0
Goiter	SD-Map	106	0.66	54	51.6

considers $|\zeta_{opt}| = 100$ too large for HotSpot on `HepStea`, the diagram would show them that a reduction to $|\zeta_{opt}| = 58$ is possible, reducing $Silh$ only slightly from 0.48 to 0.37. Also in the other direction: if $|\zeta_{opt}|$ is rather low, the diagram shows that a small increase does not change $Silh$ much; therefore, the added rules may also be important. The expert could even analyze the diagram to derive a range instead of a single value, e.g., the range where $Silh$ is above 90% of its maximum.

To assess the representativeness of the cluster proxies, we compare them with three baseline criteria that return the top k rules according to odds ratio (baseline 1), coverage (baseline 2), and WRA (baseline 3). Figure 4.4 juxtaposes the representativeness of SD-Clu and the three baselines for different numbers of rules k returned to the expert for the HotSpot algorithm on the sample `HepStea`. The plots are arranged by rule selection method (rows) and number of representative rules k returned to the expert (columns). Each graph shows the $adjCovRate$ (y-axis) of instances (x-axis) with respect to R (solid black curve). The instances are sorted by the number of rules in R they are covered by, with their respective $covRate$ shown as dots. The dotted curve represents a locally weighted scatterplot smoothing (LOWESS) of the points. Ideally, both curves are close to each other, meaning that the instances are covered by approximately the same proportion of rules in the cluster proy as the proportion of rules in general.

For all approaches, *representativeness* improves with increasing number of representative rules k. For example, *representativeness* increases from 0.87 to 0.96 for SD-Clu and $k = 10$ to $k = 50$ which means that the absolute difference between $adjCovRate$ of ζ and $covRate$ of R over all instances successively decreases. Further, for a given k, the representative rules of the baselines are less representative than SD-Clu's cluster proxy rules, e.g. 0.91, 0.92, 0.91 vs. 0.96 for $k = 50$, respectively (cf. 5th column of plot matrix in Figure 4.4).

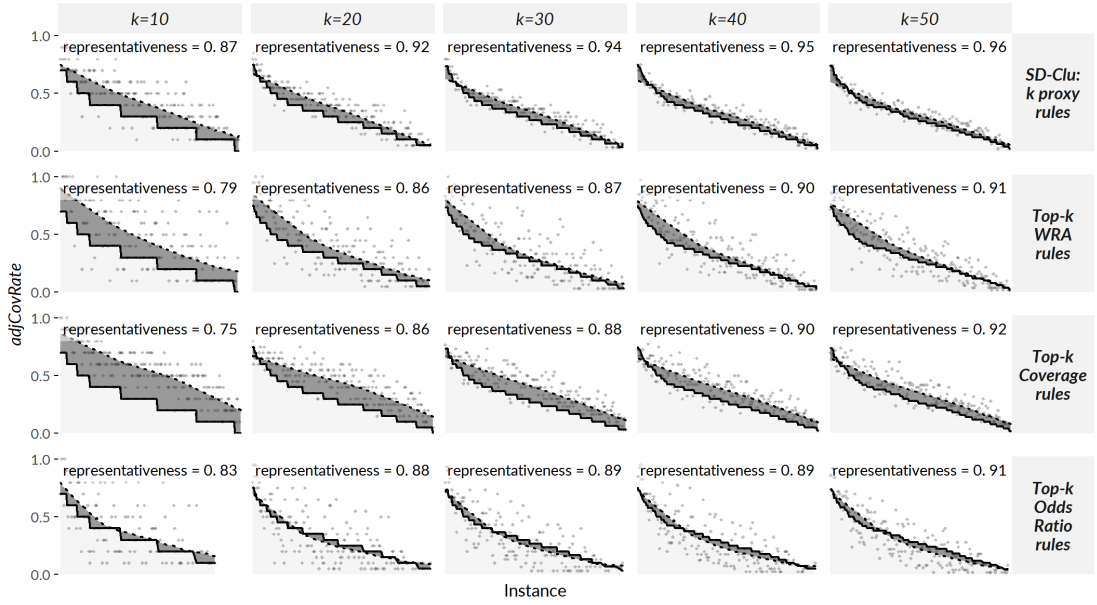

Figure 4.4: **Evaluation of** *representativeness* **on `HepStea` using the HotSpot algorithm.** *representativeness* of SD-Clu and three baseline approaches for different numbers of clusters k on the `HepStea` sample using HotSpot. Points depict an instance's *adjCovRate* with respect to the set of representative rules of the approach (row) and k (column). Instances are sorted by *covRate* with respect to R , i.e., the set of all rules) in descending order, shown by a solid black curve. A dotted curve depicts a LOWESS regression fit on the points. The similarity between solid and dotted curve illustrates *representativeness* of the top-k rules of the respective approach. This is illustrated by the dark gray area in-between solid curve and dotted curve where smaller areas are better and reflect higher *representativeness* values.

Table 4.2: **Representative rules (`HepStea`).** Proxy rules for k=5 on the `HepStea` sample learned and the positive outcome as target.

#	Antecedent of cluster proxy	Sup	Conf	Lift
HotSpot				
1	increased waist circumference = TRUE	0.72	0.37	1.69
2	hypertension = TRUE	0.60	0.33	1.54
3	age > 44 ∧ apolipoprotein A1 ≤ 1.56 g/l	0.37	0.41	1.90
4	physical health score ≤ 47.3 ∧ increased waist circumference = TRUE	0.29	0.48	2.12
5	high-sensitivity C-reactive protein > 2.8 mg/l ∧ uric acid > 246 µmol/l	0.29	0.46	2.21
SD-Map				
1	diastolic blood pressure > 79.75 mmHg ∧ hip circumference > 98.05 cm ∧ cholesterol-HDL-quotient > 3.015	0.67	0.40	1.85
2	increased waist circumference = TRUE ∧ diastolic blood pressure > 79.75 mmHg ∧ body mass index > 26.3 kg/m²	0.58	0.45	2.07
3	waist circumference > 88.15 cm ∧ diastolic blood pressure > 79.75 mmHg ∧ body mass index > 26.3 kg/m²	0.59	0.46	2.11
4	body mass index > 26.3 kg/m² ∧ alanin-aminotransferase > 0.385 µkatal/l ∧ diastolic blood pressure > 79.75 mmHg	0.55	0.48	2.20
5	body mass index > 26.3 kg/m² ∧ uric acid > 278.5 µmol/l ∧ alanin-aminotransferase > 0.385 µkatal/l ∧ treated urinary tract diseases = TRUE	0.54	0.46	2.34

4.4.1 Discussion of Findings

Tables 4.2 and 4.3 show the antecedent, support, and confidence of cluster proxy rules found by two algorithms for `HepStea` and `Goiter` with k=5. The prevalence of hepatic steatosis or goiter is significantly higher in each of the subpopulations described by these rules than in the corresponding overall population. These subpopulations are characterized by known risk factors for hepatic steatosis, such as large waist circumference and BMI, blood pressure and hypertension, advanced age, and high values in some of the medical tests (ALAT and LDL). Furthermore, apolipoprotein A1 (ApoA1), a major protein component of high-density lipoprotein (HDL) particles in plasma, was found to be associated with outcome in elderly patients (see fourth hotspot rule in Tables 4.2). Lipoprotein metabolism is considered the main process contributing to the development of fatty liver [96]. In addition, Poynard et al [143] found that patients with hepatic steatosis had higher levels of ApoA1 than patients with hepatic fibrosis, who in turn had higher levels than patients with cirrhosis. The fifth hotspot rule describes a subpopulation with elevated levels of liver high-sensitivity C-reactive protein (CRP) (approximately 0.80-quantile) and elevated levels of uric acid (approximately 0.36-quantile). Lizardi-Cervera et al. [123] reported increased levels of ultra-sensitive CRP in subjects with hepatic steatosis independent of other metabolic states. Similarly, Keenan et al. [99] found elevated uric acid levels in patients with hepatic steatosis independent of metabolic syndrome.

Similarly, the identified subpopulations for goiter (see Table 4.3) are characterized by common risk factors such as increased weight and body mass index and participants prescribed angiotensin II receptor blockers (see second HotSpot rule). Furthermore, the first HotSpot rule describes participants with intima-media thickness greater than 0.73 mm (approximately 0.80-quantile). Previous studies found associations between intima-media thickness and thyroid-stimulating hormone [171] and subclinical hypothyroidism [56, 179]. The condition of the third HotSpot rule describes the duration of an ECG phase. Jabbar et al [95] summarize in their review that pathological thyroid hormone levels increase the risk of cardiovascular disease. This

Table 4.3: **Representative rules (`Goiter`).** Proxy rules for k=5 on the `Goiter` sample learned and the positive outcome as target.

#	Antecedent of cluster proxy	Sup	Conf	Lift
HotSpot				
1	intima-media thickness > 0.73 mm	0.27	0.43	1.34
2	intake of angiotensin II receptor blocker = TRUE	0.10	0.62	1.90
3	duration of QRS complex on ECQ > 114 ms ∧ discouraged and sad mood = "never" ∧ body mass index > 26.65 kg/m²	0.06	0.87	2.68
4	education level = 8 ∧ thrombocytes < 209 Gpt/l	0.11	0.62	1.92
5	aorta descendens thickness > 2.79 mm ∧ thyroid-stimulating hormone ≤ 1.05 mU/l ∧ hemoglobin > 8.8 mmol/l	0.06	0.88	2.73
SD-Map				
1	intima media thickness > 0.55 mm ∧ proton pump inhibitors intake = FALSE ∧ age > 38.5	0.68	0.44	1.34
2	duration of ECG P wave > 111 ms ∧ age > 38.5 ∧ proton pump inhibitors intake = FALSE	0.60	0.45	1.38
3	hip circumference > 98.75 cm ∧ age > 38.5 ∧ proton pump inhibitors intake = FALSE	0.60	0.44	1.34
4	increased waist circumference = TRUE ∧ age > 38.5 ∧ proton pump inhibitors intake = FALSE	0.59	0.44	1.34
5	increased waist circumference = TRUE ∧ intima media thickness > 0.55 mm ∧ proton pump inhibitors intake = FALSE	0.51	0.46	1.41

association appears to be especially true in the elderly [47]. The fourth rule suggests that certain thrombocyte levels indicate increased thyroid volume, which confirms Erikci et al [43] who found that hypothyroid patients had higher platelet volume and platelet distribution width than a control group.

4.5 Conclusions on identifying distinct subpopulations

SD-Clu tackles the problem of high instance overlap in sets of rules generated by subgroup discovery algorithms. By limiting the number of rules time spent for rule inspection is reduced. SD-Clu nominates a representative rule from a hierarchical clustering from a large set of rules, and thus returns rules that express distinct concepts, i.e., rules that cover different sets of instances. The introduced *representativeness* measure assesses whether instances are similarly often covered by representatives as by the total rule set. SD-Clu was evaluated on two samples from an epidemiological study where an optimal set of *proxy* rules was selected that (i) contains considerably less rules than the total rule set and (ii) is more representative compared to the baseline approaches, respectively.

Chapter 5

Visual Identification of Informative Features

Brief Chapter Summary

We identify distinct tinnitus phenotypes by clustering screening data using a parameter-free algorithm that leverages the Bayesian information criterion to determine a suitable number of subgroups. We present novel radial bar and radial line visualizations to juxtapose phenotypes in a high-dimensional feature space and to explore phenotype-specific idiosyncrasies. We provide a web application with enhanced versions of these visualizations that also depict treatment effects.

This chapter is partly based on:
Uli Niemann, Petra Brueggemann, Benjamin Boecking, Matthias Rose, Myra Spiliopoulou, and Birgit Mazurek. "Phenotyping chronic tinnitus patients using self-report questionnaire data: cluster analysis and visual comparison". In: *Scientific Reports* 10.1 (2020), p. 16411. DOI: 10.1038/s41598-020-73402-8.

The supervised methods of classification and subgroup discovery described in the previous chapters show great potential for applications where there is one or a small number of well-defined target variables. However, some medical conditions have heterogeneous appearances which are not well understood yet. For example, chronic tinnitus is a complex, multi-factorial and heterogeneous symptom. Clinical assessment and selection of a suitable therapy strategies are difficult as not all patients benefit equally from treatment. Due to the large number and heterogeneity of available assessment tools, unsupervised methods like cluster analysis appear to be promising approaches to extract clinically relevant tinnitus phenotypes. Clinical acceptance of these empirical results can be further strengthened by a comprehensive visualization which intuitively illustrate major characteristics of a phenotype and differences between multiple phenotypes. In this chapter, we describe a workflow (1) to identify distinct tinnitus phenotypes by a parameter-free clustering algorithm and (2) to visualize these subgroups to explore phenotype idiosyncrasies. In Section 5.1, we describe the clinical value of patient subgroup stratification, briefly review previous approaches and list requirements for a clustering solution. We specify the features used for clustering in Section 5.2, give an overview of the clustering algorithm in Section 5.3, introduce our cluster visualization in Section 5.4 and our interactive web application

in Section 5.5. In Section 5.6, we present the phenotypes and describe their major characteristics.
The medical interpretation of our findings was developed together with tinnitus experts and is provided in Section 5.7. Further, we discuss the strength and weaknesses of our workflow in Section 5.8. Finally, we conclude the chapter with a summary in Section 5.9.

5.1 Motivation and Comparison to Related Work

Challenges for management and treatment of tinnitus are caused to a large extend by its clinical heterogeneity, which includes individual perception, risk factors, comorbidities, degrees of perceived stress and treatment response (cf. Section 2.6.4). These factors make it difficult for clinicians (a) to choose *the* treatment which is most effective for an individual patient, and (b) to design a unified treatment strategy all patients equally benefit from. The awareness of the existence of distinct patient subgroups may stimulate the development of more effective therapy modules. Since clinically relevant subgroups have not been established yet, clustering emerges as a promising approach to identify characteristic tinnitus *phenotypes* in a data-driven and hypothesis-free way.

Previous studies found subgroups of tinnitus patients with cluster analysis based on a small number of audiometric features [113], a combination of features extracted from self-reports, audiometry and psychoacoustics [178], or neuroimaging data and socio-demographics [157]. Although each of these studies provided insights in tinnitus subgroup patterns, acceptance among medical scholars may be increased by presenting the clustering results with intuitive visualizations that show individual subgroup patterns and enable the visual juxtaposition of multiple subgroups with respect to high-dimensional data.

Addressing this requirement, Schlee et al. [158] proposed a compact radar chart visualization that allows to compare the degree of health burden between individuals or subgroups based on measurements from self-report questionnaires. While their visualization could be applied to any disease domain, Schlee et al. demonstrated its efficacy showing subgroup differences with respect to measurements of tinnitus distress and associated comorbidities. However, they did not aim to visualize clustering results, but restricted themselves to pre-defined cohorts by graphically comparing female against male patients, and patients with low tinnitus frequency against patients with high tinnitus frequency.

5.2 Selection of Measurement Instruments

Discussions with tinnitus experts about the selection of measurement instruments (hereafter denoted as *features*) for clustering resulted in two main requirements: (1) features should cover the clinical heterogeneity of tinnitus to a great degree, and (2) if available, more robust compound scores should be preferred over single items from a questionnaire. From the routine questionnaire assessment battery (cf. Section 2.6.4), we selected a total of 64 features[1] from 14 questionnaires. These include all questionnaire total scores, all questionnaire subscale scores and all items of questionnaires that have neither subscales nor total scores. The features measure (general) tinnitus characteristics, physical quality of life, experiences of pain, somatic expressions, affective symptoms, tinnitus-related distress, internal resources, perceived stress, and mental quality of life.

[1] The complete list of features is provided in Appendix A.

From a total of 4,103 patients, data from 2,875 (70.1%) was incomplete and therefore excluded. The N=1,228 patients included in the final sample were only slightly, yet significantly younger than the excluded ones ($\mu_{\text{included}} = 50.0$, $\sigma_{\text{included}} = 11.9$; $\mu_{\text{excluded}} = 51.7$, $\sigma_{\text{excluded}} = 13.6$; $t(2630.8) = 4.0$, $p < 0.01$). Additionally, for 989 of the included patients patients (80.5%) post-treatment data were also available and used to visually explore treatment effect differences between clusters. Since most of the features have greater scores for higher health burden, we reversed the remaining features with greater scores for higher quality of life. A feature X is reversed as $X_{reversed} = \max(X) - X$. The asterisk suffix in a feature name (e.g. ACSA_qualityoflife*) denotes a reversed feature. Due to widely differing value ranges, each feature was standardized via z-score normalization. A feature X with expected value $E(X) = \mu$ and variance $Var(X) = \sigma^2$ is standardized into $Z = \frac{X-\mu}{\sigma}$. For Z, it holds true that $\mu = 0$ and $\sigma^2 = 1$.

5.3 Identification of Tinnitus Phenotypes using Clustering

Practical considerations of data clustering include how to set the number of clusters K. Since a ground truth is often not available, several heuristics to automatically determine K have been proposed. A popular approach is the so called "elbow" method which involves running the clustering algorithm with different values for K (cf. Section 6.2.1 for the application of the elbow method for density-based clustering). The number of clusters is plotted against cluster *compactness*. In the popular K-means algorithm, cluster compactness is quantified by total within-sum of squares (WSS), the sum of squared distances between each observation and its centroid over all clusters. As WSS or similar goodness of fit measures increase monotonically with increasing K, the idea of the elbow method is to identify the curve's characteristic "knee point" which is the first point from which adding another cluster leads only to a *minor* improvement in compactness. Because the plot is not guaranteed to exhibit such a distinctive knee point and universal compactness thresholds do not exist, this approach is sometimes impracticable. Another popular clustering evaluation measure is the Silhouette coefficient which favor clusterings where similar objects are assigned to the same cluster and dissimilar objects are assigned to different cluster.

Instead of a evaluating clustering quality post-hoc, we decided to chose an algorithm which internally identifies an appropriate number of clusters. X-means [140] is a parameter-free adaption of the popular K-means algorithm which incorporates the Bayesian information criterion [161] (BIC) to find a good trade-off between low total sum of squares and a small number of clusters.

Let $\mathcal{D}$ be the dataset with d dimensions and let D be a subset of $\mathcal{D}$, i.e., $D \subseteq \mathcal{D}$. A K-means clustering on D creates the set of clusters $\mathcal{C} = \{C_1, \dots, C_k, \dots, C_K\}$, where c_k is the centroid of cluster k, r_k is the number of objects in D assigned to C_k and p is the number of free parameters, i.e., $p = (d+1) \cdot K$. The BIC of a cluster C_k using the Schwarz criterion is calculated as

$$\text{BIC}(C_k) = \hat{l}_k(\mathcal{D}) - \frac{p_k}{2} \cdot \log|\mathcal{D}| \tag{5.1}$$

where $\hat{l}_k(\mathcal{D})$ is the log-likelihood of $\mathcal{D}$ according to C_k. The point probabilities are computed as

$$\hat{P}(x_i) = \frac{r_{(i)}}{|\mathcal{D}|} \cdot \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(\frac{1}{2\hat{\sigma}^2}||x_i - c_{(i)}||\right) \tag{5.2}$$

where the maximum likelihood estimate for the variance (under the identical spherical Gaussian assumption) is

$$\hat{\sigma}^2 = \frac{1}{|\mathcal{D}| - K} \sum_{i=1}^{|\mathcal{D}|} \left(x_i - \mu_{(i)}\right)^2. \tag{5.3}$$

The log-likelihood of $\mathcal{D}$ according to $\mathcal{C}$ is

$$l(\mathcal{D}) = \log \prod_{i=1}^{|\mathcal{D}|} P(x_i) = \sum_{i=1}^{|\mathcal{D}|} \left(\log \frac{1}{\sqrt{2\pi}\hat{\sigma}} - \frac{1}{2\sigma^2} ||x_i - c_{(i)}||^2 + \log \frac{r_{(i)}}{|\mathcal{D}|} \right). \tag{5.4}$$

The main steps of the X-means algorithm are summarized in Figure 5.1. At the start, an initial partitioning is generated by ordinary K-means with $K = K_{\text{lower}}$, where K_{lower} is a lower bound for the number of clusters. Then, each cluster is bisected; the resulting two child centroids are placed in opposite direction along a randomly chosen vector by a distance proportional to the cluster radius. For each pair of child clusters, a local K-means clustering with $K = 2$ is run. If the BIC score of the new partitioning exceeds the BIC score of the parental one, the child centroids are kept, otherwise the parent centroid is retained. The iterative steps are repeated until there is no cluster whose bisection leads to a better BIC score, or until the number of clusters exceeds an optional upper bound K_{upper}. We used the R implementation of Ishioka [93]. Since we did not aim to restrict the solution space with respect to the number of clusters, we set K_{lower}, i.e., the lowest possible value, and we did not set K_{upper}. For internal validation, we recorded the number of clusters produced by X-means on 500 bootstrap samples.

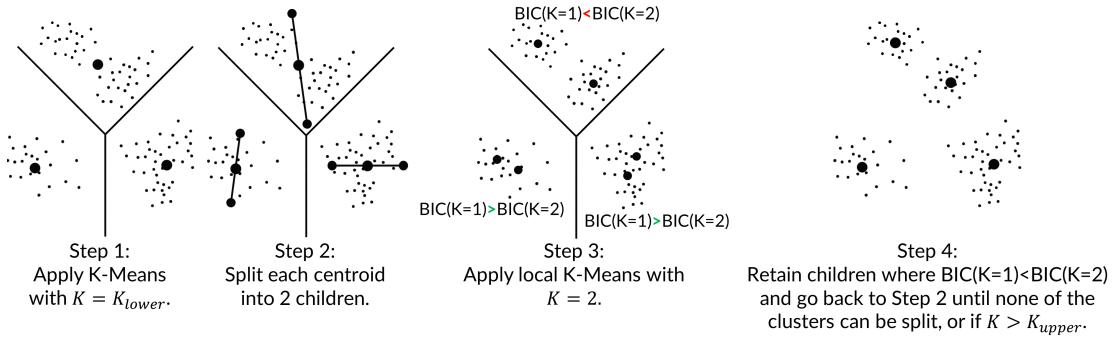

Figure 5.1: **Principal steps of X-means (simplified).** Adapted from [140].

5.4 Visualisation of Phenotypes

Visualizing clusters in high-dimensional data is challenging. Scatterplot matrices (SPLOMs) can intuitively represent the relationship between all pairs of features as a matrix of two-dimensional scatterplots [92, 102]. However, as the number of features increases, the number of scatterplots grows quadratically, leading to scalability problems such as overplotting. Several advanced visualization techniques have been proposed as a remedy, from simply adding transparency or colors to points to more sophisticated density contours, hexagon binning, layers with aggregated geometric features (Minimal Spanning Trees, Alpha Shape, Convex Hull), animation, or combinations of several techniques such as splatterplots [131]. However, SPLOMs and other traditional visualization techniques such as parallel coordinate graphs [74] are still more suitable for low-dimensional data.

Dimensionality reduction (DR) techniques are often used to project the original data onto a low-dimensional projection that allows the above visualization types to be used. Ideally, this projection preserves the most important structures of the original data, such as relative pairwise distance, clusters, outliers, and correlations. Principal Component Analysis [91] (PCA) is a seminal DR algorithm that generates linear, orthogonal combinations of the original dimensions. Each new dimension, called a principal component, contains a loading indicating how

much variability in the data it covers. Typically, the first two or three dimensions that carry the highest charges are selected for visualization. PCA is not robust to outliers and cannot capture nonlinear relationships. Multidimensional scaling [64] (MDS) is another early DR technique that emphasizes the preservation of pairwise distances, ie, Objects that are close to each other in high-dimensional space should also be close to each other in low-dimensional projected space. For complex, arbitrarily shaped structures, pairwise distances may be subject to the curse of dimensionality, leading to poor results. t-stochastic neighborhood embedding [127] (t-SNE) and Uniform Manifold Approximation and Projection [132] (UMAP) are nonlinear dimensionality reduction methods that represent a matrix of pairwise similarities. The idea is to obtain both global structures such as clusters and local structures such as distances and neighbors. Both t-SNE and UMAP can produce superior projections compared to traditional linear techniques, provided their hyperparameters are appropriately tuned. However, a shortcoming of these techniques is that the mapping to the original features cannot be quantified. Moreover, a projection cannot be applied to new observations; instead, a new projection must be recomputed. Because of their stochasticity, different runs with the same hyperparameters may yield different results. Since the semantics of the original dimensions are lost, we decided that DR methods are not suitable for this application.

Discussions with tinnitus experts led to the following cluster visualization requirements:

- keep original (interpretable) features,
- represent high-dimensional data with dozens of features,
- compactly compare multiple clusters at a glance,
- Contrast cluster characteristics with the overall patient mean.

Following these requirements, we implemented (a) a radial bar chart as a visualization of a single cluster (Figure 5.2) and (b) a radial line chart visualization for comparing multiple clusters at once 5.3. The radial bar chart is used to compare observations assigned to a single cluster with the overall population. The mean values of the features within a cluster are represented by bars arranged in a radial layout. Each bar begins at the black "zero line," which represents the feature means of the entire population, i.e., all subjects used for clustering. Because features are standardized (i.e., z-scored) prior to clustering, bars inclined to the outside represent feature averages above the population average and bars inclined to the inside represent feature averages below the population average. For example, a bar whose top is positioned at -1 characterizes a feature average within a cluster that is 1 standard deviation smaller than the population average. In addition to the combination of bar height and bar direction, within-cluster averages are also mapped to the bar color by a sequential color gradient from dark blue (lower boundary) to yellow (population average) to light red (upper boundary). Gray error lines at the top of a bar represent the within-cluster standard error. To allow quick feature localization, features were grouped into expert-defined categories, which are displayed in the inner circle along with the cluster name and the number of subjects assigned.

The radial line chart (5.3) allows a comparison of all clusters in a single display. Instead of bars, the feature mean values are represented as points. Points of the same cluster and feature category are connected by line segments. Points and line segments are colored according to their respective cluster.

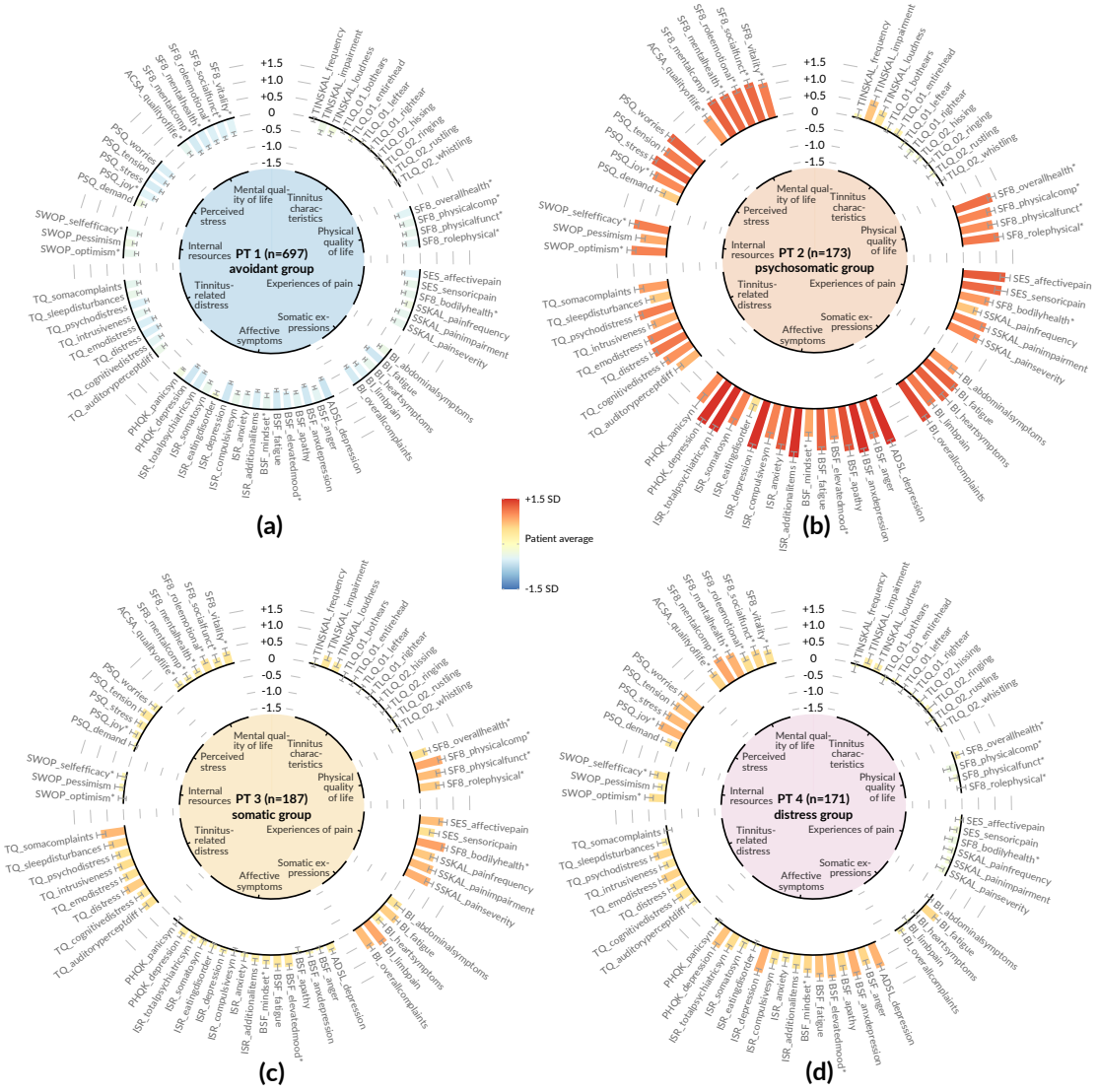

Figure 5.2: **Radial bar graphs for each of the 4 phenotypes (PT).** (a) PT 1 characterizes the subgroup with the lowest health burden. (b) PT 2 includes the most suffering subjects, in whom all mean values of psychosomatic and somatic characteristics exceed the population mean by more than 0.5 standard deviations (SD). (c) PT 3 indicates somatic indicator scores above the population mean. (d) PT 4 indicates subjects with elevated distress scores, including subjective stress and perceived quality of life. Bars are arranged in a circular layout, with the height and direction of a bar representing the within-cluster feature average and the gray line centered at the top of the bar illustrating the 95% confidence interval. The characteristics were grouped into 9 categories defined by tinnitus experts. The categories are shown inside the inner circle. See Appendix A for a feature description. Adapted from [137].

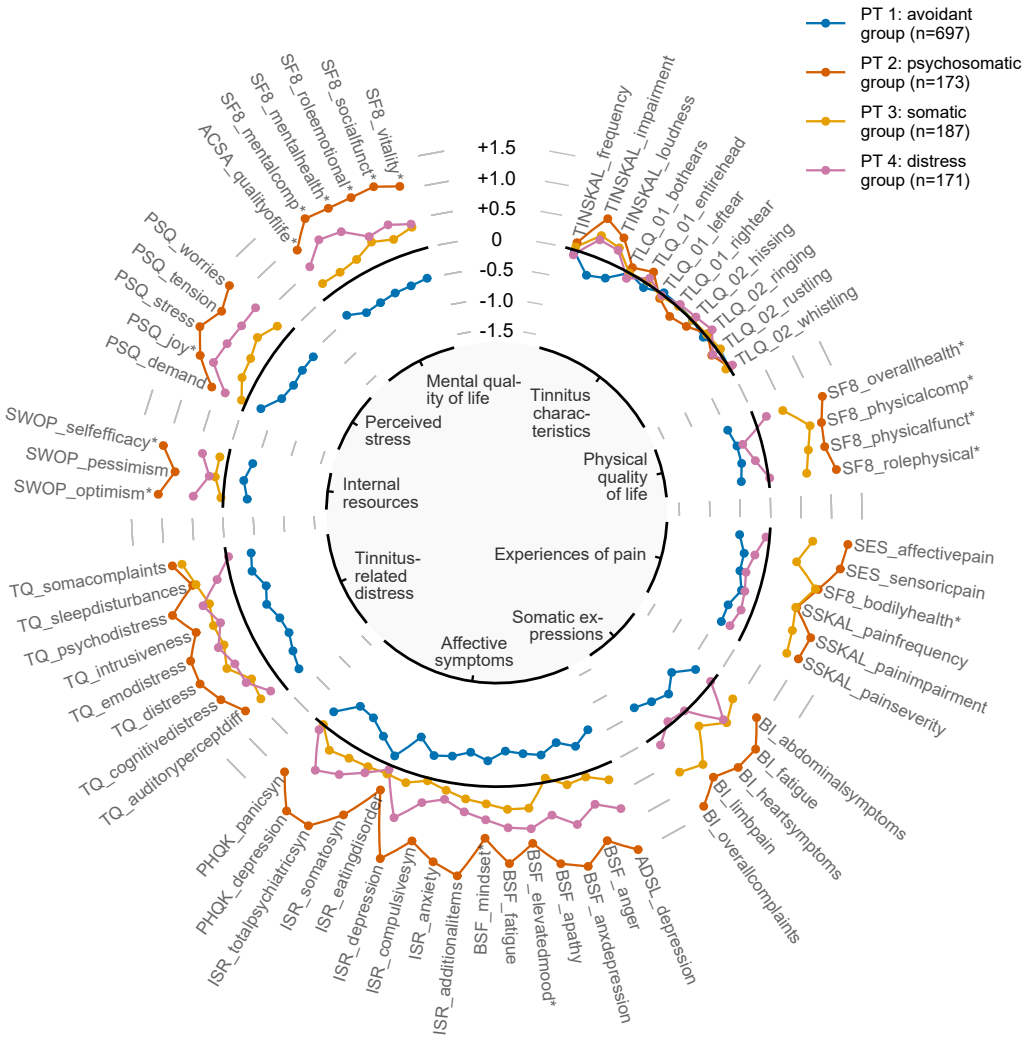

Figure 5.3: **Radial line chart for phenotypes comparison.** Points show within-phenotype feature averages. Points depicting features of the same category are connected with line segments. Points and lines are colored by cluster. See Figure 5.2 for a description of the phenotypes and Appendix A for a description of the features. Adapted from [137].

5.5 Interactive Components and Exploration of Treatment Effects

We provide an interactive demo of the cluster solutions and visualizations as a web application[2] (Figure 5.4). The following interactive components have been added: hovering over bars, lines, or feature labels in the visualizations opens tooltips with additional cluster summaries and compact feature descriptions. Clicking on a feature triggers an additional plot showing the original (unscaled) distribution of the selected feature stratified by clusters and, if selected, also after treatment. For continuous features, semi-transparent boxplots are displayed on violin plot [86] layers. For categorical traits, the points and error lines represent the category proportions and 95% confidence intervals, respectively. While clustering was performed on baseline data only, we included treatment effect indicators to allow visual detection of potential differences between clusters in treatment efficacy. To this end, we extended the radial bar chart by plotting cluster averages before treatment (T0) and after treatment (T1) as adjacent bars. The ends of a pair of bars are connected by a line with an arrowhead pointing from the T0 score to the T1 score. To highlight large treatment effects, the connecting lines and feature labels are colored according to their relative decrease from T0 to T1. Given the user-defined parameter Δ, i.e., the minimum relative magnitude of the difference between T0 and T1 considered as a change, elements were colored (a) red if the T1 score was greater than the T0 score by Δ or more, (b) green if the T0 score was less than the T0 score by Δ or less, and (c) black, respectively.

5.6 Results

Four clusters (also referred to as phenotypes hereafter) represent an optimal solution for the present data according to X-means (cf. Figure 5.5).

Phenotype 1 (PT 1) represents the largest subgroup (697 of 1,228 patients; 56.8%), characterized by substantially below-average symptom expression on tinnitus-related and more general psychosomatic symptom indices, including affective symptoms, perceived stress, tinnitus-related distress, and somatic symptoms, as well as (above-average) quality of life and internal resources (Figure 5.2 (a)). Because this group of patients is potentially more help-seeking, presents to the clinic more frequently, and wishes to participate in multimodal treatment, it can be assumed that they experience psychological distress but strive to present as unburdened as possible. Therefore, this phenotype is referred to as the *"avoidance group"*. Patients in this subgroup have comparatively high levels of education, employment, and duration of illness and psychotherapeutic treatment (Figure 5.6).

PT 2 included 173 patients (14.1%) who reported the highest emotional and somatic burden among all PTs (Figure 5.2 (b)). More specifically, PT 2 represents a patient subgroup with high psychosomatic comorbidity and is therefore referred to as the *"psychosomatic group"*. This patient subgroup shows high tinnitus burden in addition to clinically relevant impairment in all affective indices, including depression, anxiety, and perceived stress. These affective symptoms appear to be consistent with somatoform expressions of distress, including somatic symptoms. Patients in this subgroup report greatly reduced quality of life and coping behaviors, with more pessimism, less experienced self-efficacy, and optimism. Relative to the overall population, this subgroup has a higher proportion of women, patients who live alone, are unemployed, or have an overall lower educational status. Patients in this cluster also report consulting more physicians, taking more sick days, and using more psychotherapy. PT 2 patients reported that the tinnitus

[2]The demo is available at https://unmnn.de/phs/app/.

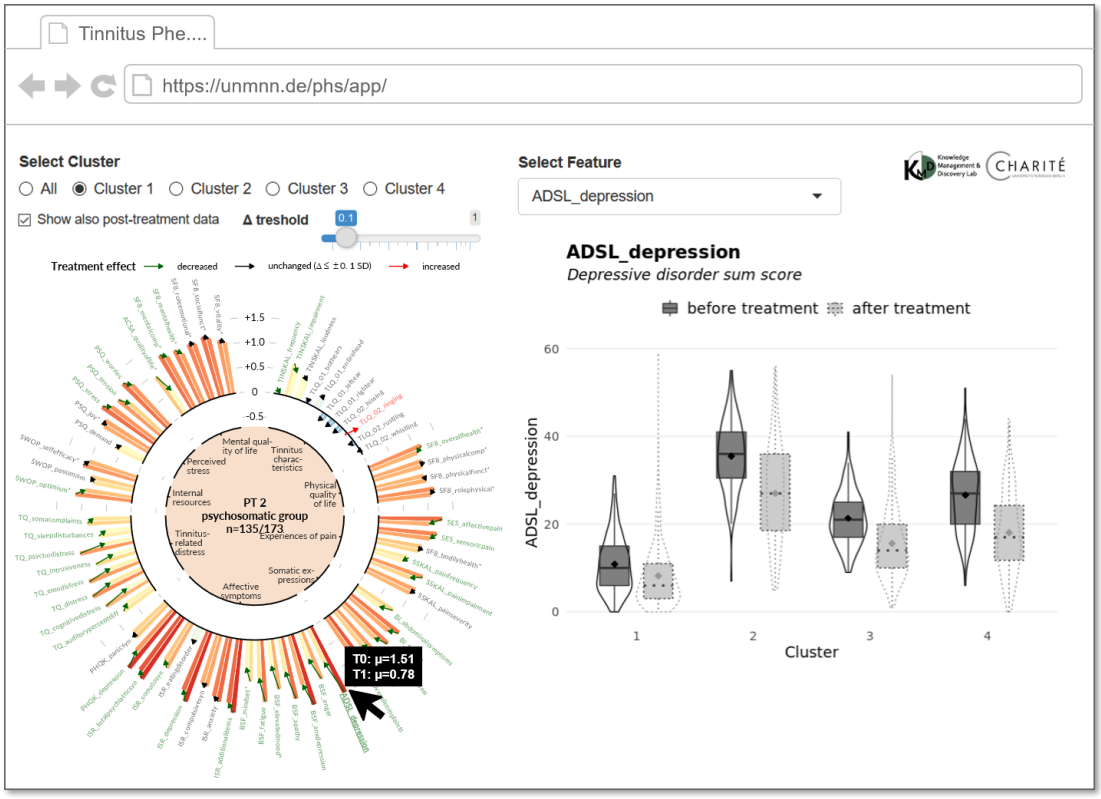

Figure 5.4: **User interface of interactive demo.** Radial bar charts were enhanced by interactive components: hovering over a bar or feature label opens a tooltip with additional cluster summaries and a compact feature description. Clicking on a feature updates the right plot showing the distribution of the selected feature stratified by cluster, and if selected, also after treatment. Continuous features are shown using semi-transparent boxplots placed on violin plot [86] layers whereas for nominal features, category proportions alongside their 95% confidence intervals are displayed as points and error lines, respectively.

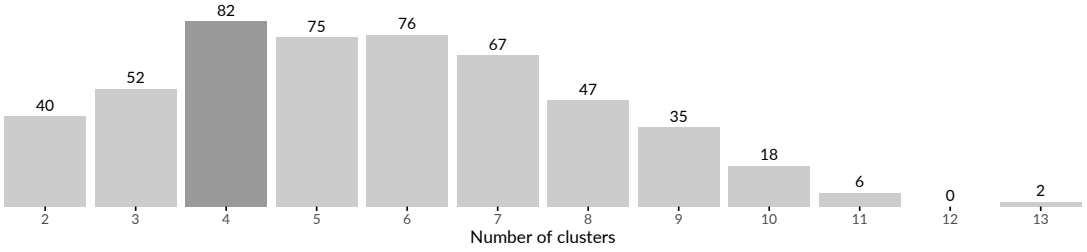

Figure 5.5: **Results of internal validation.** Bars show the distribution of the number of clusters generated by X-means for 500 bootstrap samples. The most common cluster number was 4, which occurred in 82 samples (16.4%).

noise was audible throughout the head (i.e., not unilateral) with a higher percentage than the other groups.

PT 3 contains 187 patients (15.2%) characterized by above-average scores of traits measuring somatic complaints and near-average scores for affective symptoms (Figure 5.2 (c)). Because the pain scores SF8_bodilyhealth* and SSKAL_painfrequency were similar in magnitude to PT2, this patient subgroup is referred to as the *"somatic group"*. PT 3 includes the oldest subgroup, with the largest proportion of female patients and the largest reported time since tinnitus onset.

In contrast to PT3, PT 4 (n=171; 13.9%) has above-average values for affective scores, quality-of-life components, and perceived stress (Figure 5.2 (d)), e.g., mental component summary score (SF8_mentalcomp*; 0.85) and anxious depression score (BSF_anxdepression; 0.79). Therefore, PT 4 is referred to as the *"distress group"*. PT 4 represents the youngest of the 4 subgroups (mean 47.3 years), with the largest proportion of male patients (Figure 5.6).

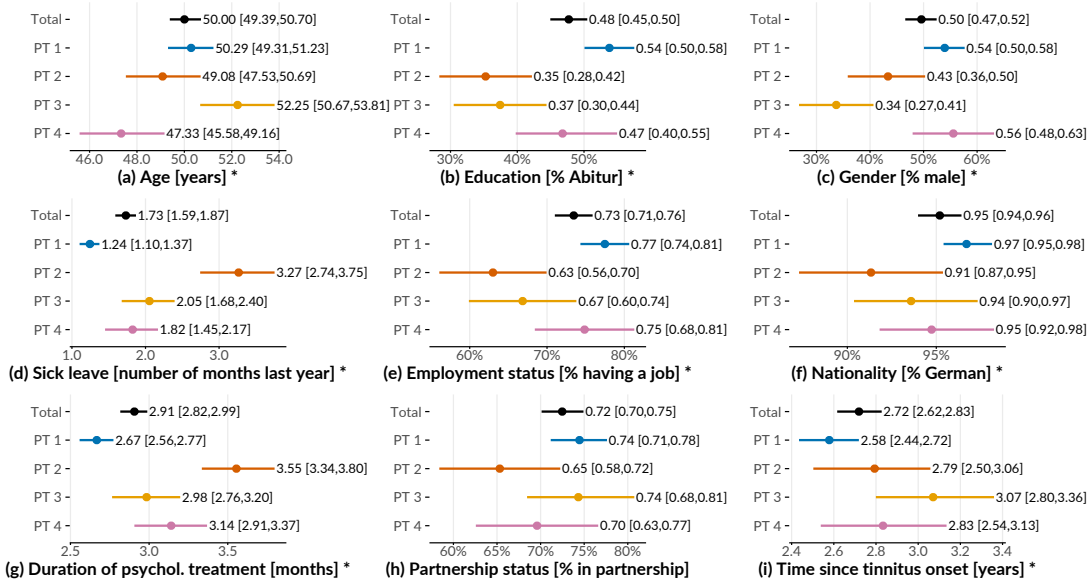

Figure 5.6: **Inter-phenotype comparison of demographic characteristics.** Summaries are given as means [95% confidence interval] for the entire population and for each of the 4 phenotypes. Confidence intervals were estimated using nonparametric Basic Bootstrap Sampling [35] with 2,000 samples each. The Kruskal-Wallis test was used for statistical comparison of differences between phenotypes for continuous features (such as age), and Pearson's chi-square test was used for categorical features (such as sex). An asterisk indicates statistical significance ($\alpha = 0.05$). Correction for multiple comparisons was not performed due to the exploratory nature of the study. Adapted from [137].

Figure 5.7 depicts the top-10 features with highest average change between T1 and T0 per cluster. For PT 1 and PT 3, BSF_elevatedmood* decreased the most, namely by 0.48 ± 0.75 and 0.65 ± 0.85 (Z units), respectively. For PT 2 and PT 4, the top-ranked features is ADSL_depression with an average difference between T_1 and T_0 of 0.73 ± 0.88 and 0.74 ± 0.83, respectively. Six out of these ten features were among the top-10 features for all clusters, including BSF_elevatedmood*, TQ_cogintivedistress, TQ_psychodistress, TQ_emodistress, TQ_distress and BSF_fatigue.

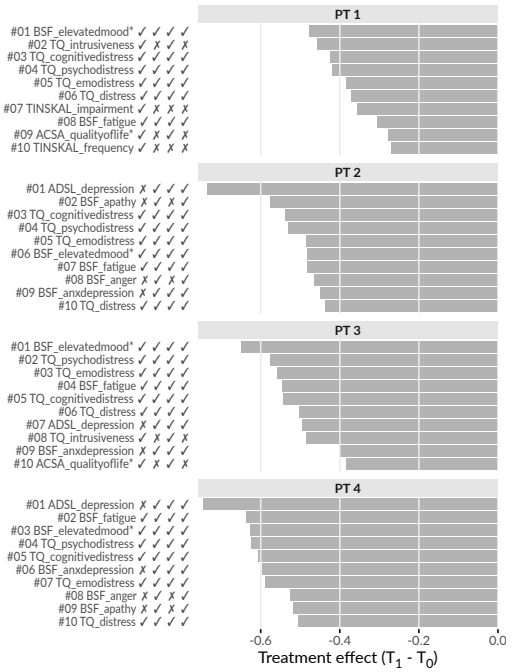

Figure 5.7: **Cluster-specific top-10 features with highest average treatment effect magnitude.** Bars depict the intra-cluster average differences between the measurements at T1 and T0 based on the standardized values. Lower values represent better treatment effectiveness. The symbols right to the feature names indicate whether the feature is among the top-10 features for the cluster at position i. For example, the character string ✗ ✓ ✗ ✓ for TQ_intrusiveness (ranked 2nd for PT 1) means that the feature is among the top-10 for PT 1 and PT 3, but not for PT 2 and PT 4.

5.7 Clinical Value

We discussed the clinical relevance of the phenotypes with three of the five tinnitus experts who co-authored the original publication [137].

PT 1 (*avoidant group*) represents more than half of the patient sample. Besides the actual tinnitus symptom, patients in this subgroup reported few other affective or psychosomatic symptoms. Because of the biased presentation of these patients (*"everything is fine if it weren't for the tinnitus"*), clinicians might easily be led to believe that assessment of possible other factors contributing to individual distress is unnecessary. However, clinical experience suggests a thorough assessment of other psychosocial stressors. The psychosocial resourcefulness of this subgroup enables patients to seek help quickly and in a solution-oriented manner. Adequate tinnitus-specific counseling and individualized (online) therapy modules that include audiological, psychological, or relaxation techniques may represent an adequate treatment strategy for this patient subgroup.

PT 2 (*psychosomatic group*) represents 15% of patients with high tinnitus distress and clinically relevant impairment across all affective indices, including depression, anxiety, and perceived stress. These affective symptoms appear to interact strongly with somatoform expressions of distress, including physical complaints and somatic symptoms. Patients in this subgroup reported greatly reduced quality of life and coping behaviors, higher pessimism, lower experienced self-efficacy, and optimism. The frequently asked question is whether increased tinnitus-related distress contributes to an increase in depression or vice versa. In this group, depressive or anxious symptoms may be considered a crucial underlying factor in overall symptom distress, and treatment must initially focus on improving mood and alleviating depression. Here, tinnitus-related distress may need to be viewed in a broader context of medical and psychological contributing factors that require patient-specific conceptualization.

PT 3 (*somatic group*) appears to represent a patient subgroup that is characterized by somatopsychic symptom expression, i.e., physical symptoms that may reflect stress and/or underlying medical conditions. To meet the needs of this patient subgroup, multimodal interventions may include a proportion of body-oriented procedures such as relaxation exercises or physiotherapy, but their effects should be interpreted in terms of both direct and indirect psychological effects (e.g., through increased well-being or affection from others).

Patients in PT 4 (*stress group*) reported higher than average perceived stress, accompanied by physical exhaustion and anxious-depressed mood. This group is more likely to include younger, employed, and male patients who reported chronic distress and may be susceptible to a burnout syndrome with subjectively reduced mental performance ("hamster wheel"), which is used to describe life situations even in the absence of tinnitus distress. Multimodal therapy should initially focus on stress regulation techniques, including relaxation or individually tailored behavior modification approaches. Similar to PT 2, which has a high psychosomatic burden, patients in PT 4 could also benefit from longer psychotherapeutic or multimodal treatment procedures (inpatient or rehabilitative).

5.8 Discussion

Although for our dataset the optimal number of clusters was four, we expect that this number may be different for a different sample of tinnitus patients even with the same clustering algorithm. Figure 5.5 shows a high variance in terms of the number of clusters returned by X-means on different bootstrap samples. Considering the only slightly lower occurrence of 5 and 6 clusters, our clustering result is certainly not set in stone.

Nonetheless, comparison of all clusters showed that some questionnaires and characteristics differed considerably between patient phenotypes. In particular, patient subgroups differed substantially in terms of coping behaviors, stress, tinnitus burden, perceived pain, discomfort, and perception of quality of life. In contrast, patients did not appear to differ with respect to localization and noise. Regarding the separability between phenotypes, the mostly high correlations between feature within the same category suggest that phenotyping is also possible with fewer questionnaires, especially since some of the questionnaires overlap semantically, e.g. PHQK_depression, ISR_depression, ADSL_depression, etc.

Without a "ground truth" and because of the different sets of available measurements, it is difficult to compare our results with those of similar studies. The greatest strength of our workflow was the inclusion of a wide range of self-report questionnaire assessments. Other studies also used audiometry [178, 113] and cardiac imaging data [157]. Nevertheless, PT 2 (psychosomatic distress group) seems to be associated with the "constant distressing tinnitus" subgroup reported by Tyler et al. [178], as the mean scores of tinnitus-related health distress were much larger than in the other subgroups. Clearly, the selection of a meaningful set of characteristics is central to the effectiveness of a cluster analysis.

The closest to our radial bar chart visualization is the radar chart proposed by Schlee et al. [158], whose solution tends to overplot when more than two subsets need to be displayed simultaneously. Therefore, we did not fill areas spanned by connected points with color to avoid that one polygon completely overlaps another. Furthermore, inferences about the radar map [158] depend heavily on the arrangement of features, since the main criterion for comparison is the shapes of the polygons. Their solution to compute an arrangement that yields areas that achieve a maximum mean area difference between subgroups and a minimum area variance within subgroups only partially solves the problem, since still only a moderate number of up to about 20 features can be represented. We chose to organize the 64 features according to expert-determined categories, e.g., quality of life, which makes it easier to find features and compare similar features.

In addition, our visualization is tinnitus-specific but can be used to display a compact summary for any condition or subset of index symptoms. Whether the visualizations will be adopted by clinicians to guide appropriate tinnitus management strategies is adopted remains to be tested. In a preliminary user study, clinicians suggested that graphical summaries of possible patient subtypes could ease the challenge of assigning an appropriate treatment strategy for specific combinations of symptom presentations.

By excluding patients who did not complete all questionnaires at T0, there may be a selection bias. Possible reasons for noncompletion include unfamiliarity with the technical equipment used to record item responses, loss of motivation due to the relatively large number of questionnaires, and collisions with other baseline studies in the laboratory. Nevertheless, the analysis of all 15 questionnaires led to insights into the contributions of these questionnaires to phenotyping, possibly allowing a reduction in the number of questionnaires. Because our results reflect only a snapshot of the patients' situation at baseline, a patient may transition from one phenotype to another at later stages of life, depending on tinnitus management. Therefore, a next step would be to investigate the effects of treatment on these phenotypes in more detail and to find out whether some patient phenotypes benefit more than others.

5.9 Conclusions

We presented a workflow to find distinct tinnitus phenotypes by clustering using an algorithm that internally determines the optimal number of clusters. We visualized the characteristics

of each phenotype and the differences between multiple phenotypes using radial bar and line graphs. We also provided a web application where phenotypes can be further explored with their treatment effects. To our knowledge, this is the first approach that combines clustering of tinnitus patients with comprehensive visualization of subgroups in high-dimensional data and interactive components for exploration of patterns and treatment effects.

Part II

EXPLOITING DYNAMICS

Chapter 6

Constructing Evolution Features to Capture Study Participant Change over Time

Brief Chapter Summary

We present a framework for cohort analysis in longitudinal cohort studies which constructs "evolution features" from latent temporal information describing the change of cohort participants over time. We show that exploiting these novel features improves the generalization performance of classification models. We report on results for SHIP and the outcome "fatty liver".

This chapter is partly based on:
Uli Niemann, Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. "Can we classify the participants of a longitudinal epidemiological study from their previous evolution?". In: *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2015, pp. 121-126. DOI: 10.1109/CBMS.2015.12.

We begin the chapter with work related to the construction of temporal representations in medical data (Section 6.1). In Section 6.2 we present our evolution feature framework including a full workflow that encompasses steps for the extraction of evolution features, dealing with class imbalance and feature selection. Subsequently, we describe our evaluation setup (Section 6.3) and present our results (Section 6.4). Finally, in Section 6.5 we conclude this chapter and give answers to the aforementioned research questions.

6.1 Motivation and Comparison to Related Work

Epidemiological studies serve as basis for the identification of risk factors associated with a medical condition. Machine learning is still rather little used in epidemiology, mainly due to the hypothesis-driven nature of their research. However, examples of machine learning applications are the identification of health failure subtypes [9] and the discovery of factors (including

biomarkers) that modulate a medical outcome [150, 180]. In longitudinal cohort studies measurements are performed in multiple study waves, hence researchers obtain access to sequences of recordings. In the context of machine learning, extracting and leveraging the inherent temporal information from these sequences may increase model performance and thus, the understanding of the medical condition of interest.

In clinical applications temporal information is often exploited, predominantly for the analysis of patient records. For example, Pechenizkiy et al. [139] analyzed streams of recordings to predict patient rehospitalization after a health failure treatment. Sun et al. [170] computed the similarity between streams of patients from patient monitoring data. Combi et al. [33] reported on streams of life signals, in particular on the temporal analysis of the timestamped medical records of hospital patients. However, participants of an epidemiological, population-based study are not hospital patients – they are a random sample of the studied population, often with skewed class distribution. In a *longitudinal* study of this kind, recordings for the same cohort member are made at each moment. Hielscher et al. [83] presented a feature engineering approach to extract temporal information from multiple, but few patient recordings in a longitudinal epidemiological study. First, for each assessment clusters of feature-value sequences associated with the target variable are found. Afterwards, original and sequence features are used in conjunction for classification. Hielscher et al. [83] showed that classification performance increases when features with temporal information are incorporated into the feature space. Instead of modeling the individual change of measurement values, our approach involves deriving multivariate change descriptors. Patient evolution with clustering was studied by Siddiqui et al. [164] who proposed a method that predicts the evolution of a patient from timestamped data by clustering them on similarity and predicting cluster movement in the multi-dimensional space. However, the patient data considered in [164] are labeled at each moment.

Our workflow combines labeled and unlabeled timestamped data from a longitudinal study to improve classification performance on skewed data. As the target variable, we study the multifactorial disorder hepatic steatosis (fatty liver) on a sample of participants from the longitudinal population-based "Study of Health in Pomerania" (SHIP) [188], see Section 2.6.1. For the SHIP cohort, the assessments (interviews, medical tests, etc.) were recorded in three *moments* (SHIP-0, SHIP-1 and SHIP-2), that are ca. 5 years apart. Temporal information is often used when analyzing patient data in a hospital, but there the time granularity is different. For example, in an intensive care unit, timestamped data are collected at a fast pace, i.e., every minute or even every second. In contrast, the participants of a longitudinal epidemiological study are monitored for a period of months or even years. This implies that measurements of the same assessment in an epidemiological dataset are few and possibly far apart. The large time span between two consecutive recordings complicates the application of methods designed for data that arrive with a higher frequency. For example, a participant may exhibit alcohol abuse or become pregnant, stop smoking and start again, take antibiotics that affect the liver, or experience other lifestyle changes that turn the medical recordings taken 5 years ago irrelevant for the learning of the participant's current health state. Another patient may have no changes in lifestyle and no illnesses, so their past data reflect only aging.

A further challenge is that only the recordings in SHIP-2 are labeled. A reliable estimate of the fat accumulation in the liver was computed from magnetic resonance tomography images. In SHIP-0, MRT was unavailable. Instead, liver fat accumulation was derived from ultrasound – a procedure with lower clinical accuracy. In SHIP-1, the calculation was omitted altogether. As a consequence, for a given SHIP participant a class label is available in SHIP-2, no label in SHIP-1 and a partially reliable indicator in SHIP-0. Since hepatic steatosis is a reversible disorder, label imputation – by means of a growth model [165] – is not possible; participant evolution must be learned with only one moment with labeled data.

We address these challenges as follows. First, we group study participants at each moment on

similarity, thus building clusters of cohort members that have similar recordings at one of the three moments. Then, we connect the clusters across time, thus capturing the transition of each cluster from one study wave to the next. These transitions reflect the *evolution of subgroups*, not individuals. Hence, next to the single labeled recording per cohort participant, we also exploit the earlier, unlabeled recordings, the description of the cluster they are assigned to and information on how the clusters evolve over time. We show that this new, augmented dataset, combining labeled and unlabeled data on individuals and on subgroups, improves classification and delivers additional insights on some factors associated to hepatic steatosis.

6.2 Evolution Features

We leverage latent temporal information of a longitudinal cohort study dataset by extracting informative features based on the individual change of participants and the transition of their respective clusters over time. For this purpose, we exploit the similarity among participants at each moment, as surrogate to the labels which are not available in the first two moments, assuming that similar participants evolve similarly. We call these new features *"evolution features"*. Our approach is illustrated in Figure 6.1 (a). We monitor the individual change of participants across the study waves, trace the change of the clusters separately, extract new features (from labeled *and* unlabeled data) and augment the original data space with our new descriptors of change. The complete classification workflow is depicted in Figure 6.1 (b).

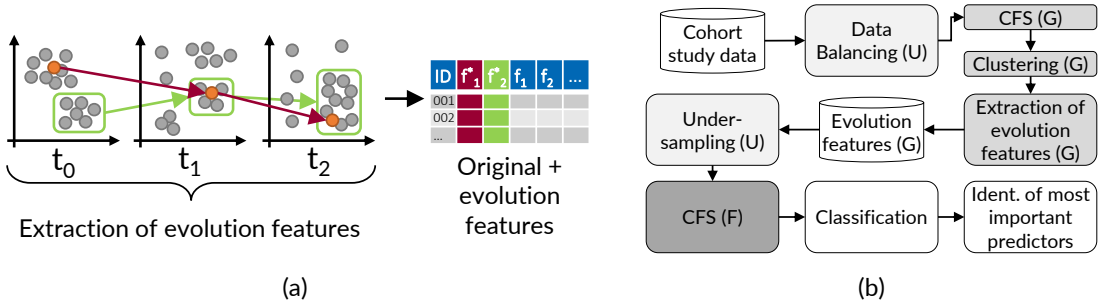

Figure 6.1: **Concept of evolution feature extraction for classification performance improvement**. (a) Clustering of longitudinal cohort data and subsequent generation of evolution features from the change of individuals (red) and whole clusters (green). (b) Overview of the classification workflow. **TODO: explain parentheses**

In the following, we describe the clustering of study participants (Section 6.2.1), the generation of evolution features (Section 6.2.2), our strategy of undersampling the majority class to balance class distribution (Section 6.2.3) and our feature selection strategy to extract a subset of *informative* features as input for classification (Section 6.2.4).

6.2.1 Clustering

For clustering, we prefer density-based clustering over partitional algorithms (like K-means), because our data contain extreme cases, the clusters may be arbitrarily shaped and of different sizes, and we cannot determine their number in advance. At each moment t, we run the DBSCAN [44] algorithm to cluster the set $Z(t)$ of recordings of all cohort members observed at t. For participant x, $v(x, f, t)$ denotes the value of x for feature $f \in$ feature-set F at t, and $obs(x, F, t)$ the set of all feature recordings for x at t (cf. notation in Table 6.1).

Table 6.1: **Symbols and basic functions.**

Term	Description
x	a study participant from cohort X
t	a study moment, one of $\{t, \dots, t_T\}$
t	a study moment, one of $\{t, \dots, t_T\}$
f	a feature from the set of all features F
$v(x, f, t)$	the value of x for feature f at moment t
$obs(x, F, t)$	all measurements for x at t, i.e., $\{v(x, f, t) \mid f \in F\}$
$Z(t)$	all observations at t, i.e., $\{obs(x, F, t) \mid x \in X\}$
$c(x, t)$	cluster membership of x at t; if x is an outlier at t, then $c(x, t)$ is empty
$d(x, z, t)$	distance between x and z at t (cf. Eq. 6.1)
$kNN(x, k, t)$	the set of k nearest neighbors of x at t
$centr(x, t)$	centroid of $c(x, t)$

Distance function. For the distance between participants x, z at t, we use the *adjusted heterogeneous Euclidean overlap metric* [83, 195], which weights the difference between two values x, z for feature f by the feature's information gain $G(f)$, scaled to the largest observed value G^*, defined as:

$$d(x, z, t) = \sqrt{\sum_{f \in F} \left(\frac{G(f)}{G^*} \cdot \delta\left(v(x, f, t), v(z, f, t)\right) \right)^2}. \tag{6.1}$$

For continuous features, $\delta(a, b)$ is the min-max-scaled difference between the values a, b, i.e., $(a - b)/(\max(f) - \min(f))$. For nominal features, $\delta(a, b)$ is 0 if $a = b$, and 1 otherwise.

DBSCAN Parameter setting. DBSCAN relies on two parameters: the radius *eps* of the neighborhood around a data point, and the minimum number *minPts* of neighbors for a point to be a core point. We use the "elbow" heuristic of [44] which determines a suitable *eps* value for a given *minPts* value, and is illustrated in Figure 6.2. More specifically, we define a parameter k, and compute for each $x \in Z(t)$ the distance k-dist(x, k) to its k-th nearest neighbor. We sort these distances, draw the k-dist(x, k) graph g and span the line l connecting the smallest k-dist() value to the largest one. Then, we set *eps* to the k-dist value with maximum distance between g and l.

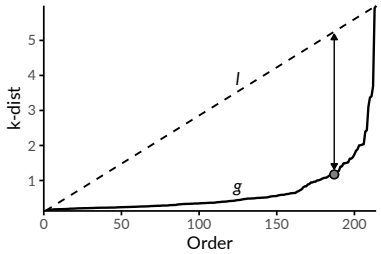

Figure 6.2: **Setting *eps* based on the k-dist graph for a given *minPts*.** The k-dist graph g depicts the sorted distances to the points' k-th next neighbors. A suitable eps_{opt} can be identified at the position with maximum distance between the k-dist and the line l that connects the first and the last point of g. For DBSCAN clustering with $minPts$ = k, eps_{opt}, points with k-dist $\leq$ eps_{opt} will become core points, else border or noise points.

6.2.2 Constructing Evolution Features

Table 6.2 provides a description of all evolution features. For each cohort member x and moment t, we record the cluster containing x (feature 1 in Table 6.2), (2) the distance of x to this cluster's centroid, (3) the fraction of positively labeled participants among the k nearest neighbors of x, (4) the (graph-based) cohesion [172] and (5) Silhouette coefficient [172] of x, and (6) the (graph-based) separation [172] of x to cohort participants outside this cluster. We compute the difference of the cohesion, silhouette and separation values from t to all later moments $\{t' \in T | t' > t\}$ (7-9), and also check how much the values of these metrics change as x moves from $c(x,t)$ to $c(x,t')$ (10-12). We record whether x is an outlier, i.e., a DBSCAN noise point at some moment (13). For t and $\{t' \in T | t' > t\}$, we compute the fraction of cohort members who are in the same cluster as x in t and t' (14), and the fraction of common k nearest neighbors (15), and the change of the distance between x and its centroid at t, from t to t' (16). We further record changes in the sequence of values for a feature, including real (17), absolute (18) and relative (19) differences between the values at two moments. We measure how a cluster shrinks/grows from t to t' (20), and how much its members move (on average) closer or far apart from their previous positions (21-23). In that way, we extract information about the evolution of the participants, distinguishing among those that evolve smoothly and those that switch among clusters. We transfer this information to evolution features, thus enriching the feature space with information from the unlabeled moments.

6.2.3 Undersampling

For imbalanced data, feature selection and classification are often biased in favor of the majority class [121]. To avoid this problem, prior to the application of CFS we undersample the majority class and generate a balanced data set to select features informative with respect to all classes.

6.2.4 Feature Selection

We use the feature selection method of Hielscher et al. [84] as follows. First, we invoke correlation-based feature selection [71] (CFS), which builds up a feature set by iteratively inserting to it the feature that adds most "merit" to it. The merit M_F of a feature set F is computed by calculating the information gain for each pair of features in F (lower gain corresponds to low correlation and is thus preferred) and for each feature in F towards the target variable (higher gain is better). Continuous features are first discretized with the entropy-based method of Hielscher et al. [46]. We discretize only for feature selection; for clustering and classification we use the original values.

As shown in Figure 6.1, we perform feature selection twice. The first time, we consider only features recorded in all moments. This is essential for evolution tracing: we can only compute distances between objects in clusters located in the same topological space. After generating the evolution features, we build up the complete set of features, also considering those not recorded in each moment. On this set we perform feature selection again, to discard unpredictive (original or evolution) features. This final feature set is then used for classification.

6.3 Evaluation Setup

We evaluate our workflow with 10-fold cross-validation on four off-the-shelf classification algorithms: random forest [21] (RF), C4.5 decision tree [146], Naïve Bayes (NB) and k-nearest

Table 6.2: **Overview of extracted features.** The first group of features (I) comprises the cluster membership and aggregated distance information for each participant and for each moment; feature group II is on changes in the participant's position (in the hyperspace) relative to the cluster and to its closest neighbors; feature group III captures changes in the values of the participant's recordings; feature group IV refers to changes in the clusters.

#	Name	Description
I: Features for participant x at each moment		
1	`Cluster_t`	Cluster ID of x at t
2	`dist_To_Centroid_t`	distance of x to the centroid of cluster $c(x,t)$, denoted as $\widehat{c(x,t)}$
3	`fraction_Of_POS_kNN_`k`_t`	fraction of the k nearest neighbors of x at t from the positive class
4-6	`a_t := a(x,t,c(x,t))`	where a is one of cohesion, silhouette, separation [172]; cohesion_t is the cohesion of x at t w.r.t. the members of $c(x,t)$ – and similarly for silhouette and for separation
II: Evolution features linked to each participant x		
7-9	`a_Delta_t_t'`	difference $a_t - a_{t'}$ for a as above
10-12	`a_Movement_t_t'`	difference $a(x,t,c(x,t)) - a(x,t',c(x,t))$ for a as above, referring to the same cluster $c(x,t)$ at two moments t,t'; at t', $x \in c(x,t')$, which does not need to be the same as $c(x,t)$
13	`was_becomes_Outlier_t_t'`	4-valued flag on whether x was outlier in both t,t', only in t, only in t' or in neithert,t'
14	`same_Cluster_t_t'`	$c(x,t) \cap c(x,t') \ \{x\}$: set of cohort members that are in the same cluster as x in t and in t'
15	`same_kNN_`k`_t_t'`	$kNN(x,k,t) \cap kNN(x,k,t')$, i.e., the set of cohort members who are among the k nearest neighbors of x in both t and in t'; they do not need to be in the same cluster as x
16	`shift_To_Old_Centroid_t_t'`	difference $d(x,\widehat{c(x,t)},t') - (x,\widehat{c(x,t)},t)$
III: Evolution features associated w. the value of each original feature f for participant x		
17-19	`A_Diff_f_t_t'`	difference between $v(x,f,t)$ and $v(x,f,t')$ for feature f, where A is either real difference, absolute difference or relative difference to $v(x,f,t)$
IV: Evolution features linked to a whole cluster c		
20	`smaller_Cluster_Fraction_t_t'`	difference of the size of cluster c at t' with respect to the size it had at t; c is matched to the clusters of t on the basis of member overlap
21-23	`movement_d_t_t'`	distance d between the locations of the members of cluster c at t and their locations at t', where d is one of Euclidean distance, HEOM distance (cf. Eq. 6.1, Cosine similarity

Table 6.3: **Workflow variants.** `UFG` is the complete workflow.

Workflow components	Under-sampling	Feature selection	Evolution features
`UFG`	✓	✓	✓
`UF-`	✓	✓	✗
`-FG`	✗	✓	✓
`-F-`	✗	✓	✗
`--G`	✗	✗	✓
`Baseline`	✗	✗	✗

neighbor (kNN). Next, we compare the generalization performance for each algorithm when it is used alone (*baseline* variant) vs. when incorporated into our workflow (*worflow-enhanced* variant). Further, we study the impact of different combinations of the three workflow components *undersampling* (`U`), *feature selection* (`F`) and *incorporation of generated evolution features* (`G`). As shown in Table 6.3, `Baseline` simply invokes the classification algorithm; we use the classification algorithms which achieves the highest F-measure scores. The variant `U-G` performs undersampling and uses the generated evolution features for classification. Since we undersample only for feature selection and then build the classification models on the original dataset, so `U-G` is identical to `--G` and `U--` is identical to the `Baseline` variant, so we omit to explicityl list `U-G` and `U--`.

The main parameter is k which is the number of neighbors of a data point: we set $minPts = k$ and use k to derive the values of the DBSCAN parameter *eps* (cf. Section 6.2.1) and of the parameters for the features `same_kNN_`k`_t_1_t_2` and `fraction_Of_POS_kNN_`k`_t` (cf. Table 6.2). Further, the number of nearest neighbors for the k-NN classification algorithm is also set to k. We vary k to measure its impact on classification performance.

Following the findings in Chapters **??** and **??** on the differences between female and male participants with respect to the outcome, we run the experiments on the whole dataset ($Partition_{all}$), and on the partitions of female ($Partition_f$) and of male ($Partition_m$) participants. Finally, we list the most important features found in $Partition_{all}$ and its two subsets.

6.4 Results

Figure 6.3 shows sensitivity (left), specificity (center) and F-measure (right) for the simple classifiers (gray curves) and their workflow-enhanced counterparts (same line style, colored), for different k. Overall, each workflow-enhanced variant outperforms its simple counterpart with respect to sensitivity and F-measure, and outperforms or performs slightly worse with respect to specificity. The workflow-enhanced Naive Bayes performs best with respect to sensitivity for any k and best for $k = 31$. Decision trees exhibit highest F-measure, with improvements on sensitivity and F-Measure compared to its simple variant, albeit specificity being sligthly worse; improvements are less for large k. Random Forests benefit the most from our workflow, with an absolute improvement in F-measure of over 30% (green vs gray "+" curves in right part of Figure 6.3). One explanation for the rather poor sensitivity of the simple RF variant is the large number of trees (100) learned on data samples containing very few positive examples, and RF could be trapped by the many majority class examples. This is consistent with the specificity curve (almost straight line around 95%) of simple RF, while the F-measure is slightly above 40%. Our workflow improves RF sensitivity (63%) and F-measure (65%), while specificity remains high (90%). Overall, the impact of k on the three measures is limited for all algorithms except of

the workflow-enhanced and the baseline k-NN which is naturally affected stronger by the value of k than any other algorithm. Therefore, the workflow-enhanced variants outperform their simple counterparts in terms of sensitivity and F-measure. For some algorithms, our workflow prevents overfitting of the negative class.

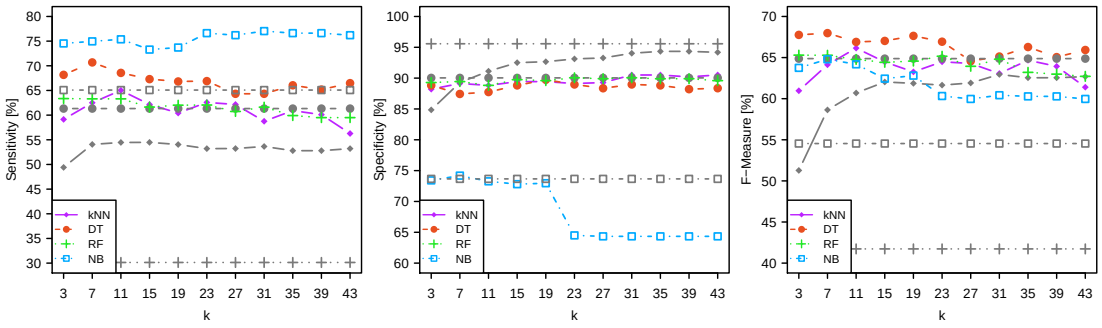

Figure 6.3: **Comparison of classification performance between workflow and baseline.** Sensitivity (true positive rate), specificity (false positive rate) and F-measure scores of different classifiers when varying the number k of neighbors to a cohort member which impacts the clustering result. For each classifier two performance curves are shown: a colored one for the workflow-enhanced version and a gray one for the baseline counterpart. Higher values are better for all measures. From [135].

The workflow component-specific analysis results in Figure 6.4 show that our complete workflow `UFG` and the variants `UF-` and `--G` outperform `---` in sensitivity and F-measure. The variants `-F-` and `-FG` perform well only regarding specificity, which suggests that feature selection may not be fruitful without undersampling for datasets with class imbalance.

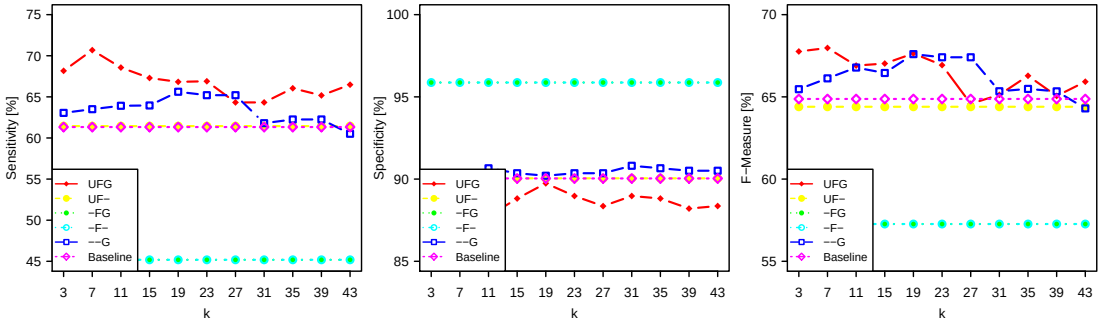

Figure 6.4: **Comparison of classification performance for workflow components.** Sensitivity, specificity and F-measure scores for each workflow variant and the baseline using decision tree for learning. From [135].

Important features. The performance of the workflow variants which include feature selection indicates that a small number of features is sufficient for class separation. Hereafter, for each partition we report on the evolution features selected for classification and among the top-15 features according to information gain. Figure 6.5 shows that for $\text{Partition}_{\text{All}}$ 3 out of these 15 features are generated evolution features. The boxplots (a) and (c) in Figure 6.5 refer to differences between values recorded in two moments. The feature `separationDelta_g_1_2` measures the difference in cluster separation for each participant based on the cluster assignment in moment 1 and 2, and corresponds to entry #9 in Table 6.2. Participants of the positive exhibit

a higher median in `separationDelta_g_1_2` than participants of the negative class indicating that clusters harboring mostly positive participants cover larger, more sparse areas.
The feature `relative_Difference_som_huef_g_0_1` (#19) quantifies the difference in a participant's hip circumference between SHIP-0 and SHIP-1, relative to the value in SHIP-0. While on average study participants from both classes lose weight when they grow older, negative participants reduce more weight compared to positive participants (cf. Figure 6.5 (c)), which in general reflects differences in life styles. The mosaic chart in Figure 6.5 (b) for feature `fraction_of_Positives_kNN_1_g_2` (#3) indicates that the *nearest neighbor* of a participant with fatty liver is more likely to also exhibit the disorder than it is for a non- fatty liver participant.

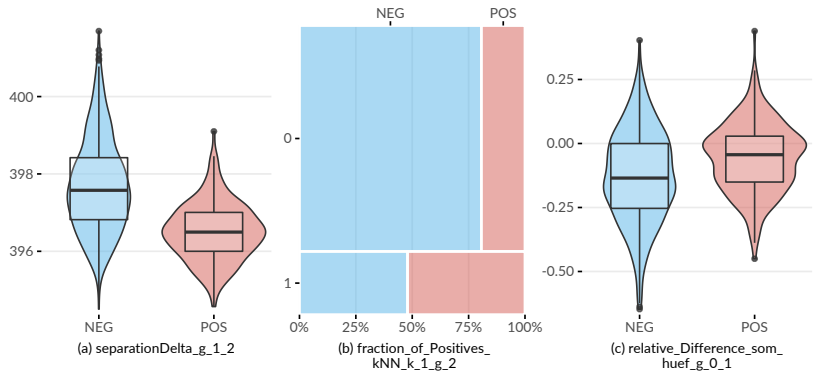

Figure 6.5: **Visualization of selected evolution features which contribute most to class separation for the whole dataset Partition$_{\mathbf{All}}$.**

For Partition$_{\mathrm{F}}$, 5 out of the top-15 features are evolution features, cf. Figure 6.6. Compared with female participants without the disorder, female subjects with fatty liver exhibit a larger distance to the centroid of their cluster in SHIP-1 (#2), a lower silhouette coefficient in SHIP-1 (#5), a higher difference in waist circumference between SHIP-0 and SHIP-2 (#19), and a lower relative difference in serum triglycerides concentration between SHIP-0 and SHIP-1 (#19).

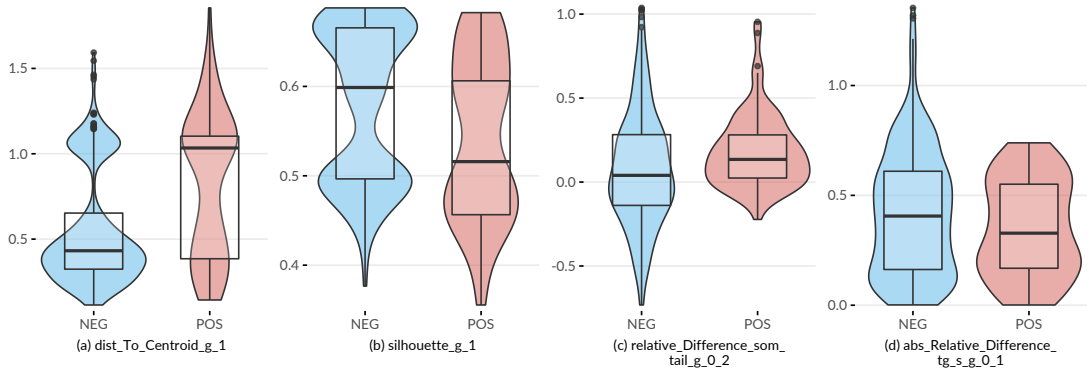

Figure 6.6: **Visualization of selected evolution features which contribute most to class separation for Partition$_{\mathbf{F}}$.**

For Partition$_{\mathrm{M}}$, 2 out of the top-15 features are evolution features (Figure 6.7), including the relative difference in waist circumference between SHIP-1 and SHIP-2 (#19) and difference in separation between SHIP-0 and SHIP-1 (#6). For both features, patients exhibiting the disorder have greater values.

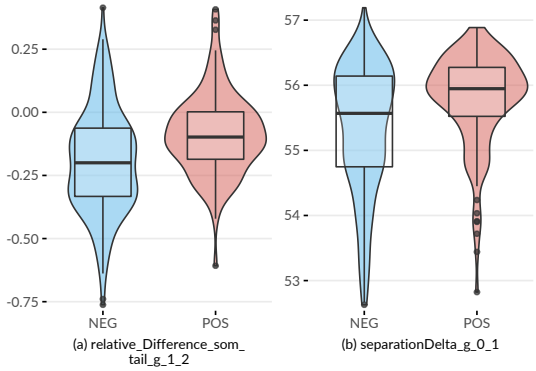

Figure 6.7: **Visualization of selected evolution features which contribute most to class separation for Partition$_{\mathbf{M}}$.**

6.5 Conclusions from Exploiting Study Participant Evolution

TODO:LINK TO RESEARCH QUESTIONS TODO: say that in the original submission, we also report on static features

We proposed a workflow for the the classification of longitudinal cohort study data which exploits inherent temporal information by clustering the cohort participants at each moment, linking the clusters and tracing participant evolution over the moments of the study. From the clusters and their transitions we extract *evolution features*, which are added to the feature space and subsequently used for classification. The workflow improves the generalization performance with respect to sensitivity and F-measure scores. The generated evolution features contribute to this improvement, even when they are used alone and without undersampling of the skewed data. We show that the *change* of the values of somatographic variables and cluster quality indices over time are predictive.

Chapter 7

Feature Extraction From Short Temporal Sequences for Clustering

Brief Chapter Summary

We present an approach to build representations from short temporal sequences via clustering by the example of pressure- and posture-dependent plantar temperature and pressure in patients with diabetic foot syndrome.

This chapter is partly based on:

- Uli Niemann, Myra Spiliopoulou, Fred Samland, Thorsten Szczepanski, Jens Grützner, Antao Ming, Juliane Kellersmann, Jan Malanowski, Silke Klose, and Peter R. Mertens. "Learning Pressure Patterns for Patients with Diabetic Foot Syndrome". In: *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2016, pp. 54–59. DOI: 10.1109/CBMS.2016.31.
- Uli Niemann, Myra Spiliopoulou, Thorsten Szczepanski, Fred Samland, Jens Grützner, Dominik Senk, Antao Ming, Juliane Kellersmann, Jan Malanowski, Silke Klose, and Peter R. Mertens. "Comparative Clustering of Plantar Pressure Distributions in Diabetics with Polyneuropathy May Be Applied to Reveal Inappropriate Biomechanical Stress". In: *PLOS ONE* 11.8 (2016), pp. 1–12. DOI: 10.1371/journal.pone.0161326.
- Uli Niemann, Myra Spiliopoulou, Jan Malanowski, Juliane Kellersmann, Thorsten Szczepanski, Silke Klose, Eirini Dedonaki, Isabell Walter, Antao Ming, and Peter R. Mertens. "Plantar temperatures in stance position: A comparative study with healthy volunteers and diabetes patients diagnosed with sensoric neuropathy". In: *EBioMedicine* 54 (2020), p. 102712. DOI: 10.1016/j.ebiom.2020.102712.

to be written

Part III

POST-MINING FOR INTERPRETATION

Chapter 8

Post-Hoc Interpretation of Classification Models

Brief Chapter Summary

We present a machine learning workflow that combines classification of high-dimensional medical data and model explanation using post-hoc interpretation methods. To this end, we use Shapely value explanations (SHAP), LASSO coefficients, and partial dependency graphs. Our approach provides statistics and visualizations representing global feature importance, instance-individual feature importance, and subpopulation-specific feature importance, all of which help illuminate complex black-box machine learning models. We report our results on three applications: (i) tinnitus-related distress in tinnitus patients, (ii) depressivity in tinnitus patients, and (iii) rupture risk in intracranial aneurysms.

This chapter is partly based on:

- Uli Niemann, Philipp Berg, Annika Niemann, Oliver Beuing, Bernhard Preim, Myra Spiliopoulou, and Sylvia Saalfeld. "Rupture Status Classification of Intracranial Aneurysms Using Morphological Parameters". In: *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2018, pp. 48-53. DOI: 10.1109/CBMS.2018.00016.
- Uli Niemann, Benjamin Boecking, Petra Brueggemann, Wilhelm Mebus, Birgit Mazurek, and Myra Spiliopoulou. "Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables". In: *PLOS ONE* 15.1 (2020), pp. 1-18. DOI: 10.1371/journal.pone.0228037.
- Uli Niemann, Petra Brueggemann, Benjamin Boecking, Birgit Mazurek, and Myra Spiliopoulou. "Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and sociodemographics". In: *Scientific Reports* 10.1 (2020), p.4664. DOI: 10.1038/s41598-020-61593-z.

In medical applications, understanding and clearly communicating the results of a machine

learning model is critical to deriving actionable knowledge that can ultimately be used to improve disease prevention, diagnosis, and treatment. Obtaining results that are easily understood by both data scientists and medical experts helps formulate new hypotheses regarding the relationship between potential risk or protective factors and the target; the significance of these relationships can be tested in follow-up studies. Current state-of-the-art machine learning algorithms produce models with superior performance compared to simpler but interpretable models, such as decision trees, rule lists, or linear regression fits. However, because these opaque *black boxes* involve many complex feature interactions or decisions, some of which are nonlinear, it is often difficult to explain them in an understandable way. Arising from the need to provide understandable insights into otherwise opaque models, the *interpretability* community of machine learning has gained traction with the goal of resolving the dilemma of choosing between moderately accurate but interpretable models and highly accurate but opaque black-box models.

In this chapter, we describe a comprehensive data analysis workflow for high-dimensional medical data that includes classification, feature elimination, and post-learning analysis steps in addition to application-specific preprocessing steps. A variety of learners are used for classification, from simple interpretable models to complex black boxes. For the best classifier, we explore different post-hoc interpretation methods to derive model-, observation-, and subpopulation-level insights. We report our results and evaluate our approach based via experiments on three applications.

This chapter is organized as follows. In Section 8.1, we describe reasons for using interpretable machine learning methods and provide methodological underpinnings of a selection of pioneering methods. Subsequently, we present the components of our mining workflow in Section 8.2, which includes correlational analysis, image preprocessing, feature selection, classification of high-dimensional medical data, hyperparameter tuning, model evaluation and our approach of putting interpretation methods into use. In Section 8.3, we report our findings on three applications: prediction of (i) tinnitus-related distress and (ii) depression after treatment in tinnitus patients, as well as (iii) rupture status classification in intracranial aneurysms. We discuss these results in Section 8.4 and conclude the chapter in Section 8.5.

8.1 Motivation and Methodological Underpinnings

Current state-of-the-art machine learning algorithms, such as gradient boosting [53] for tabular data and deep learning [63] for unstructured data (images, videos, audio recordings), are widely used to support medical decision-making. These methods produce models which typically achieve better predictive performance than simpler models such as decision trees, rule lists, or linear regression fits. However, they are also more complex, making it more difficult to understand why a prediction was made. Thus, practitioners may face the dilemma of choosing either an opaque black-box model with high predictive power or a simple, less accurate model that can at least be explained to the domain expert. Especially in high-risk domains such as healthcare, where misconceptions can have serious consequences, the ability to explain the reasoning of a model is a highly desirable, if not essential, property of any decision-support system [66, 133].

As a result, methods that explain the predictions of complex machine learning models have attracted increasing attention in recent years [25, 1]. Existing methods are classified according to different criteria [133]. For example, a distinction is made between *intrinsically interpretable models* and *post-hoc explanations*. The former often entails limiting model complexity by choosing algorithms that produce transparent models, such as decision trees or linear regression models. Decision trees, for example, can be intuitively visualized with node-link diagrams. Features in split conditions near the tree root generally have a higher impact on predictions than features occurring at lower tree levels or within leaf nodes. Quantitative measures calculate the overall

importance of a feature by the decrease in impurity or variance in nodes where the feature occurs compared to parent nodes [98]. Furthermore, the data partitions created by a decision tree can be described by understandable conditions such as "body mass index > 30," and the decision paths from root to leaf nodes provide insights into feature interactions. In addition, they can be used for contrasting predictions for individual instances, e.g., by considering alternative feature values and their effects on model prediction (*"If the patient had a body mass index of 25 instead of 30, what difference in terms of prediction would that have."*). Disadvantages are that decision trees are not able to capture linear, non-axis-parallel relationships between predictors and response, and they can be unstable with respect to small changes in training data [75]. Therefore, they may be unsuitable for very complex learning tasks.

If a more sophisticated model is trained instead, post-hoc methods can be applied to examine the model after training. The output of these methods can be *feature summary statistics*, *model internals*, *individual observations*, and *feature summary visualization* [133]. In general, feature summary statistics are individual scores that express the overall importance of a feature to model prediction or the strength of the feature's interaction with the other features. Examples of model internals are the coefficients of a linear model or weight vectors of a neural network. Individual observations can describe representatives (or prototypes) of observation subgroups for which the model provides consistent predictions for all subgroup members. Individual observations can also be used to provide counterfactual explanations, e.g., to determine the minimum change that will cause the model to predict a different class for a particular observation of interest.

Partial dependence plots. Visualizations of feature summaries typically depict trends in the relationship between a subset of features and the predicted response, often in the form of curves or surface plots. The *partial dependence plot* [53] (PDP) is a widely used tool for visually depicting the marginal effect of one or more predictors on the predicted response of a model. As an example, the PDP in Figure 8.1 (a) shows a roughly S-shaped relationship between the values of the predictor and the values of the response estimated by the model.

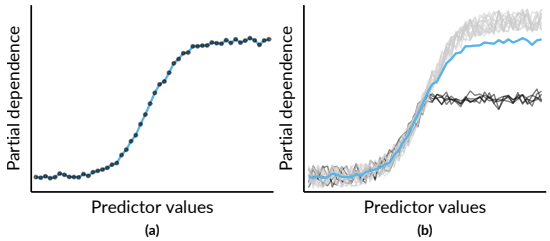

Figure 8.1: **Illustrations of a partial dependence plot (PDP) and an individual conditional expectation (ICE) plot on artificial data.** (a) PDP for a predictor on an artificial dataset. Points represent a sample of the predictor distribution. (b) PDP augmented with ICE curves. There are 2 distinct subsets of observations for which the PD is different in the upper half of the predictor distribution.

Let ζ be the classification model (or any general function that returns a single real value) and let $F = Q \cup R$ be the total set of features, where Q is the chosen subset of features and R is the complement subset. The partial dependence PD of a model ζ on Q can be represented as

$$PD(Q) = \mathbb{E}_R\left[\zeta(X)\right] = \int \zeta(Q, R) p_R(R)\, dR. \tag{8.1}$$

Here, $p_R(R)$ is the marginal probability density of R, i.e., $p_R(R) = \int p(X)\, dQ$, where $p(X)$ is the joint density of dataset X. When this complement marginal density $p_R(R)$ is estimated from the training data, PD can be approximated as

$$PD(Q) = \frac{1}{N} \sum_{i=1}^{N} \zeta(Q, R_i) \tag{8.2}$$

where R_i are the actual values of the complementary features for observation i, and N is the total number of observations in the training data. The cardinality of Q is usually chosen to be either equal to 1 or 2. The results are visualized as a line chart (if $|Q| = 1$) or a contour chart (if $|Q| = 2$). In practice, a random sample is often drawn from Q to reduce computation time.

Because averaging across all observations removes information about variability, PD curves can obscure the potentially distinct observation subgroups with substantially different effects between predictors and model output. As a remedy, Goldstein et al. [62] proposed *individual conditional expectation* (ICE) plots showing a curve for each observation. Figure 8.1 (b) illustrates an example where there are a small number of observations (black curves) that differ from the rest because their PD is constant for the second half of the predictor distribution.

LIME. Another criterion for distinguishing model interpretation methods is whether their explanations are *global* or *local*, i.e., whether the explanations apply to all observations or only to one or a small number of selected observations. *Local Interpretable Model-Agnostic Explanations* [151] (LIME) is a popular local post-hoc interpretation method. The main assumption of LIME is that a complex model is linear on a local scale [151]. Thus, to explain the predictions of a black-box model for a particular observation of interest i, LIME generates a surrogate model that is intrinsically interpretable and whose predictions are similar to the predictions of the black-box model in the *proximity* of i. The main ideas of LIME are shown in Figure 8.2. Figure 8.2 (a) shows the decision boundary of a black-box model. Since the form of the nonlinear decision boundary is quite complex, the model and its predictions cannot be explained in simple terms. LIME attempts to approximate the behavior of the black-box model by creating a linear surrogate model that performs well, especially near a user-selected instance of interest. To this end, a *perturbed* training set is created by repeatedly randomly changing the values of the instance of interest. Figure 8.2 (b) shows the instance of interest and the perturbed instances, where the glyph size represents the proximity to the instance of interest.

A linear *surrogate model* is then trained on this dataset, with observation weights proportional to their distance from the instance of interest. In Figure 8.2 (b), the decision boundary of the surrogate model is shown by the dashed line. Finally, model internals are displayed to the user as an explanation, such as the coefficients of a logistic regression model. Figure 8.2 (c) shows a feature importance ranking, where the bar height represents the model coefficient of a feature. While LIME provides intuitive interpretations and is applicable to both tabular and non-tabular data, there are several design decisions to make and hyperparameters to tune, including neighborhood kernel and width, surrogate model family, feature selection mechanism, number of features considered for the surrogate model, among others. The stability of the results of LIME has been questioned [3, 184].

Model-specific interpretation methods are limited to specific model families, while *model-agnostic* interpretation methods can be applied to any type of model. Model-specific methods are based on model internals and are widely used for neural networks [156], e.g. layered relevance propagation [10], which explicitly uses the layered structure of a neural network to infer explanations. In contrast, model-agnostic methods are decoupled from the actual learning process and do not have access to algorithmic internals. Since they only consider the output of the models, i.e., the predictions, most model-agnostic methods are also post-hoc. For example, LIME is a representative of a model-agnostic, post-hoc interpretability method.

SHAP. Closely related to LIME is the *Shapley Additive Explanations* [125] (SHAP) framework, which derives additive feature attributions to the predictions of a model. SHAP is based on

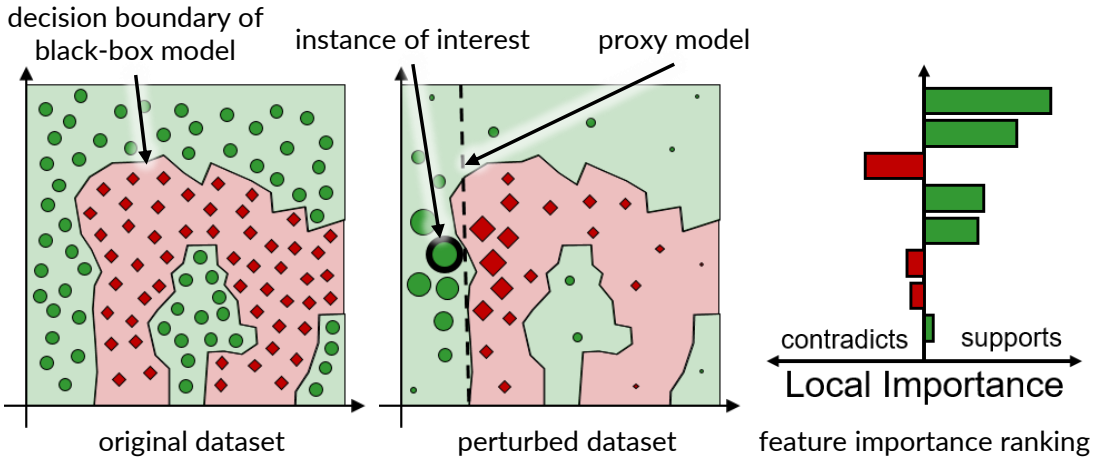

Figure 8.2: **Illustration of LIME's main ideas.** (a) A data set with a two-class problem, represented as a two-dimensional scatterplot for simplicity. The nonlinear decision boundary of a black-box model cannot be easily explained. (b) LIME aims to approximate the predictions of a black-box model for the vicinity of an instance of interest by an intrinsically interpretable model, such as a logistic regression model. The dashed line shows the linear decision boundary of this surrogate model. (c) A feature importance ranking can be derived from the model coefficients.

Shapley values [122, 169, 163], originally developed for game theory. The term "additive" denotes that for a given observation, the model output should be equal to the sum of the attributions of each feature. More specifically, for observation x, the model output $\zeta(x)$ is

$$\zeta(x) = \phi(\zeta, x)_0 + \sum_{j=1}^{M} \phi(\zeta, x)_j \tag{8.3}$$

where $\phi(\zeta, x)_0 = E(\zeta(x))$ is the expected value of the model over the training data, $\phi(\zeta, x)_j$ is the attribution of feature j for x, and M is the total number of features. Then, for each combination of feature j and observation x, the Shapely value ϕ represents the impact of each predictor being added, aggregated by a weighted average over all possible feature subsets $S \subseteq S_{all}$:

$$\phi_j(x) = \sum_{S \subseteq S_{all} \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} \left(\zeta_{S \cup j}(x) - \zeta_S(x)\right). \tag{8.4}$$

The SHAP feature importance estimates offer several practical properties. First, the sum of the feature attributions for an observation is equal to the difference between the average prediction of the model and the actual prediction for that observation (*local accuracy*). Second, if a feature is more important in one model than in another, regardless of which other features are also present, then the importance attributed to that feature should also be higher (*symmetry / monotonicity*). Third, if a feature value is missing, the associated feature importance should be 0 (*missingness*). Several approaches have been proposed to reduce the complexity of Shapley value estimation from exponential to polynomial time, including KernelSHAP [125], which works on any model type, and TreeShap [126] for tree-based models.

Feature Selection. In the context of predictive modeling, feature selection (FS) methods generally aim to reduce the number of predictor features to either (a) maximize model performance or (b) affect model performance as little as possible. Some modeling families are sensitive to

predictors that are irrelevant to the target feature, such as support vector machines [18] and neural networks [181, 63]. Others, such as linear and logistic regression models, are susceptible to correlated predictors. Often domain experts require intrinsically interpretable models, which requires eliminating predictors that do not contribute substantially to model performance.

Traditionally, FS methods are broadly classified into three categories: embedded, filter, and wrapper [68]. Embedded FS refers to internal mechanisms of modeling algorithms that evaluate the usefulness of features. Examples of such algorithms include tree- and rule-based models [147, 108], regularization methods such as Least Absolute Shrinkage and Selection Operator [52] (LASSO) and Ridge [88] regression. Filtering methods rank predictors only once based on some measure of importance, e.g., correlation with target feature. Popular examples include correlation-based feature selection [72] and Relief [101]. Wrapper methods rank and refine candidate feature subsets through an iterative search driven by model performance. Examples include sequential forward search, recursive backward elimination, and genetic search [28].

8.2 Overview of the Mining Workflow

In this section, we describe the components of our mining workflow (Figure 8.3). We apply the workflow on three classification tasks: prediction of (i) tinnitus-related distress and (ii) depression after treatment in tinnitus patients, as well as (iii) rupture status classification in intracranial aneurysms. We refer to these tasks as CHA-Tinnitus, CHA-Depression and AneurD, hereafter.

TODO: COMPLETE

For the CHA dataset, we selected patients with complete data for each of the two classification tasks. For AneurD, we segmented the aneurysms from the raw image data, performed automated centerline and neck curve extraction, and generated the morphological features. We performed correlation analysis to identify relevant correlations between predictors, correlations between predictors and response, and significant differences in correlation between predictors and response between T0 and T1. We embedded model training in an iterative feature elimination wrapper that retained predictors identified as important to the model. We selected the best overall model based on AUC and used post-hoc interpretation methods to identify predictors with the highest attribution to model prediction on a global, subpopulation and observation level.

8.2.1 Preprocessing of Raw Image Data

For AneurD, it was necessary to extract the morphological features first.

Segmentation and neck curve extraction. Aneurysms and vessels were segmented using a threshold-based approach [59] from digital subtraction data reconstructed from 3D rotational angiography images. Subsequently, the centerline of the vessel was extracted using the Vascular Modeling Toolkit (VMTK, vmtk.org) [4]. Subsequently, the plane separating the aneurysm from its parent vessel was determined using the automatic ostium detection of Saalfeld et al. [154].

Morphological feature extraction. For each 3D surface mesh, we obtained the neck curve, the dome point D, and the two base points B_1 and B_2. As described in [154], B_1 and B_2 were approximated as points on the centerline with largest distance where the rays from B_1 and B_2 to D do not intersect the surface mesh. Figure 8.4 illustrates the extracted parameters, where H_{max}, W_{max}, H_{ortho}, W_{ortho}, and D_{max} (see Figure (a)) describe the aneurysm shape [37, 118]. The angle parameters α, β, and γ (Figure 8.4 (b)) were extracted based on B_1, B_2, and D,

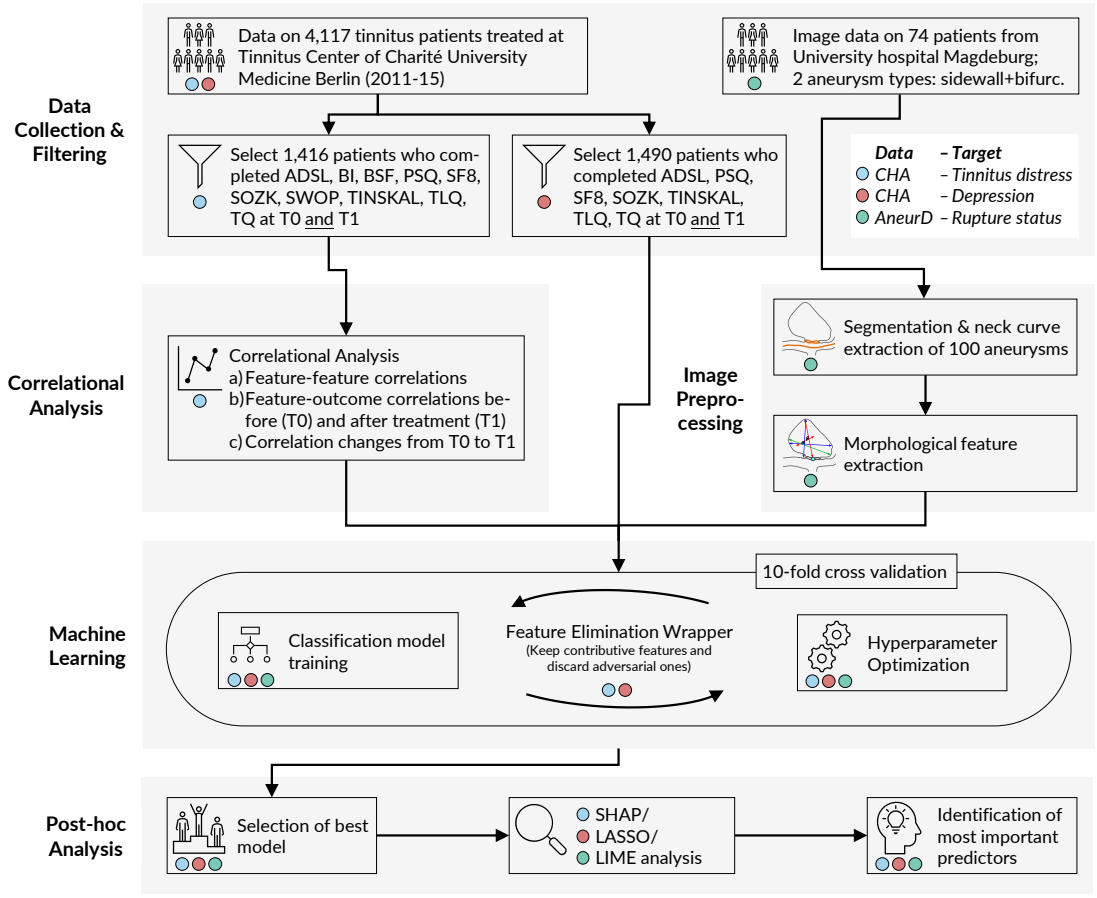

Figure 8.3: **The mining workflow.**

respectively. The absolute difference between α and β is denoted as $\Delta_{\alpha\beta}$. By separating the aneurysm from its parent vessel by the neck curve, we were able to derive the surface area A_A and volume V_A of the aneurysm (Figure 8.4 (c)). We provide two variants for the surface area of the ostium, A_{O1} and A_{O2} (see Figure 8.4 (d)). A_{O1} is the area of the ostium, i.e., the area of the triangulated ostium surface resulting from the connection of the neck curve points with their centroid C_{NC}, and A_{O2} denotes the area of the neck curve when projected into a plane (cf. [154]). Therefore, A_{O2} was extracted as a parameter comparable to other studies that often use a cutting plane to determine the ostium. For highly lobulated aneurysms, our method achieves a local optimum and considers only one of the many dome points. Although the estimated positions of B_1 and B_2 may vary slightly, neck curve detection is still performed and morphological parameters are calculated. Table (tab:09-morphological-features) provides an overview of all extracted morphological features.

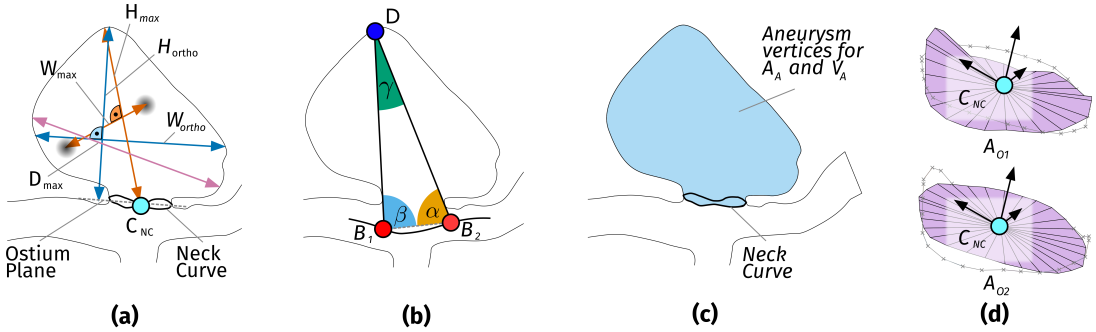

Figure 8.4: **Illustration of the extracted morphological features.** (a) Features that describe aneurysm width, height, and diameter. (b) The angles α, β and γ are extracted from the base points B_1, B_2 and the dome point D. (c) After separating the aneurysm from its parent vessel via the neck curve, the area A_A and volume V_A are computed. (d) The area of the ostium A_{O1} and the area of the projected ostium A_{O2} are extracted after estimating the center of the neck curve C_{NC}.

8.2.2 Correlational Analysis

For CHA tinnitus, we calculate pairwise Spearman correlation between the predictors as part of exploratory data analysis. Using agglomerative hierarchical clustering with complete linkage, we arrange predictors in a correlation heat map so that potential subgroups of predictors with similar intra-group correlations and similar inter-group correlations can be visually identified. Furthermore, we calculate the median correlation between the response and the predictors from the same questionnaire at T0 and T1 to obtain potential candidate predictors important in the later modeling step. In addition, we identify predictors with the highest absolute correlation with respect to the response at T0 and T1, respectively. Finally, we examined predictors whose correlation with the TQ distress score differed most between T0 and T1.

8.2.3 Classification Algorithms

In order not to be limited to a particular classification algorithm, but to create a model with the highest possible predictive power, we examined a total of eleven classifiers:

- Least absolute shrinkage and selection operator [52] (*LASSO*) and *Ridge* [88] are extensions of ordinary least squares (OLS) regression that perform feature selection and regulariza-

Table 8.1: **Overview of morphological features extracted for AneurD**.

#	Feature	Description
1	A_A	area of the aneurysm (without the ostium) [mm^2]
2	V_A	volume of the aneurysm [mm^3]
3	A_{O1}	area of the ostium (variant 1) [mm^2]
4	A_{O2}	area of the ostium (variant 2) [mm^2]
5	D_{max}	max. diameter of the aneurysm [mm]
6	H_{max}	max. height of the aneurysm [mm]
7	W_{max}	max. width of the aneurysm perpendicular to H_{max} [mm]
8	H_{ortho}	height of the aneurysm approximated as length of the ray perpendicular to the ostium plane starting from C_{NC} [mm]
9	W_{ortho}	max. width parallel to the projected ostium plane [mm]
10	N_{max}	max. NC diameter, i.e., the max. possible distance between two NC points [mm]
11	N_{avg}	average NC diameter, i.e., the mean distance between C_{NC} and the NC points [mm]
12	AR_1	aspect ratio (variant 1): H_{ortho}/N_{max}
13	AR_2	aspect ratio (variant 2): H_{ortho}/N_{avg}
14	V_{CH}	volume of the convex hull of the aneurysm vertices [mm^3]
15	A_{CH}	area of the convex hull of the aneurysm vertices [mm^2]
16	EI	ellipticity index $EI = 1 - (18\pi)^{\frac{1}{3}} V_{CH}^{\frac{2}{3}} / A_{CH}$
17	NSI	non-sphericity index, i.e., $NSI = 1 - (18\pi)^{\frac{1}{3}} V^{\frac{2}{3}} / A$
18	UI	undulation index. $UI = 1 - \frac{V}{V_{CH}}$
19	α	min. of $\measuredangle DB_1B_2$ and $\measuredangle DB_2B_1$ [deg]
20	β	max. of $\measuredangle DB_1B_2$ and $\measuredangle DB_2B_1$ [deg]
21	γ	angle at D, i.e., $\measuredangle B_1DB_2$ $[deg]$
22	$\Delta_{\alpha\beta}$	abs. difference between α and β [deg]

tion to improve both predictive performance and interpretability. For a dataset with n observations, p predictor features and a target y, the objective of LASSO and Ridge is to solve

$$\underset{\beta}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)\right)^2}_{\text{Residual Sum of Squares}} + \alpha\lambda \underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{L1 Penalty}} + (1-\alpha)\lambda \underbrace{\sum_{j=1}^{p}\beta_j^2}_{\text{L2 Penalty}} \tag{8.5}$$

where β are the to be determined model coefficients, and λ is a tuning hyperparameter that controls the amount of regularization. LASSO uses the L1 norm penalty term, i.e., $\alpha = 1$, which shrinks the absolute values of the coefficients, often forcing some of them to be exactly equal to 0. Ridge uses the L2 norm penalty term, i.e., $\alpha = 0$, which shrinks the coefficient magnitudes. In general, LASSO performs better than Ridge when there is a relatively small number of predictors with substantial coefficients and the remaining predictors have coefficients that are close or equal to zero. Ridge performs better in settings where the response depends on many predictors, each of them with approximately equal importance. From the perspective of interpretability, LASSO has the advantage of producing sparser models by reducing the values of some of the predictors' coefficients to exactly zero.

- Partial least squares is another derivative of OLS regression which first performs a projection to extract latent variables which capture as much of the variability among the predictors as possible while modeling the response well. A linear regression is then fit on a preferably small number of latent features from this projection. We use the generalized partial least squares (*GPLS*) implementation from Ding and Gentleman [40].
- A support vector machine (*SVM*) [18] learns linear or nonlinear decision boundaries in the feature space to separate the classes. The decision boundary is represented by the training observations that are most difficult to classify, i.e., the *support vectors*. The goal is to find the *maximum margin hyperplane* which is the separating hyperplane with the maximum margin to the support vectors. In case a linear decision boundary does not exist, nonlinear SVM approaches can be used which apply the so called *kernel trick* to transform the original feature space into a new, higher-dimensional space in which a linear hyperplane can be found to separate the classes.
- An artifical neural network (*NNET*) consists of a structure of nodes that are connected with each other by directed edges. Each node performs a basic unit of computation. Nodes are supplied by data values that are passed over via incoming edges from other nodes. Each edge holds a weight that controls the impact on the node it forwards values to. The main goal of a NNET is to adjust the weights of the edges such that the relationship between predictors and response in the underlying data is represented. Neural networks extract new useful features from the original predictors that are relevant for classification. By combining interconnected nodes to complex predictive features, NNETs are capable of extracting more classification-relevant feature sets compared to expert-driven feature engineering or by dimension reduction techniques. NNETs have undergone widespread adoption in the last decade and led to various success stories in computer vision and natural language processing [63]. We used a feed-forward NNET with one intermediary layer (*hidden unit*) [182].
- Weighted k-nearest neighbor [77] (*WKNN*) is a variant of KNN classification. To classify an observation with unknown response value, the k *nearest* training observations are identified and the modus of their response values are taken as prediction. Proximity between observations is quantified by a distance measure such as Euclidean distance. Whereas in ordinary KNN all neighbors have equal influence on the prediction, weighted KNN takes into account the actual distance magnitudes. As a result, WKNN assigns weights to training observations that are inversely proportional to their distance from the observation

being classified.

- A Naïve Bayes classifier (*NB*) uses Bayes' theorem to calculate class membership probabilities. The naive property refers to the assumption of class-conditional independence among the predictors, which is employed to reduce computational complexity and to obtain more reliable class-conditional probability estimates.
- Classification and regression trees [20] (CART), C5.0 [147], random forests [21] (RF) and gradient boosted trees (GBT) [53] are tree-based models. Algorithm from this model family partition the predictor space into a set of non-overlapping hyperrectangles based on combinations of predictor-value conditions, such as "IF age $>$ 52 & body-mass index $<$ 25". A new observation is classified based on the majority class of training data associated with the hyperrectangle to which it belongs. Random forests and gradient boosted trees are ensembles of several different decision trees, with each tree casting a vote for the final prediction. In a random forest, the base trees are created independently. In a a gradient boosted model, the base trees are constructed and added to the composite model in a way that any new tree reduces the error of the current set of trees.

8.2.4 Classifier Evaluation and Hyperparameter Tuning

We use 10-fold stratified cross-validation (CV) for classifier evaluation. In k-fold CV, the observations are split into k disjunct partitions. Each partition serves once as test set for a model trained on the remainder of the partitions. The k performance estimates are aggregated to obtain an overall performance score. We performed a grid search for hyperparameter selection (cf. Table 8.2). Because the three applications have dichotomous responses with different skew, accuracy might be inappropriate to estimate generalization performance. Instead, we used the area under the receiver operating characteristic curve (AUC) as performance measure. A receiver operating characteristic curve (ROC) shows the relationship between sensitivity (true positive rate (TPR)) and false positive rate (FPR) for a binary classifier. The area under the ROC curve (AUC) takes values from 0 (0% TPR, 100% FPR) to 1 (100% TPR, 0%FPR). A higher AUC suggests that the classifier is better at separating the classes.

Table 8.2: **Overview of hyperparameter tuning grid.** All classifiers were implemented with the statistical programming language R [148] using the package `mlr` [17], which provides a uniform interface to the listed machine learning algorithms from other R packages. A grid search was used to tune the hyperparameters using area under the ROC curve (AUC) as the evaluation measure. The table provides an overview of each classifier, including the R package used, the tuned hyperparameters and their value ranges. All other hyperparameters were set to default values. * = {linear, polynomial, radial, sigmoid}

Algorithm (R package)	**Parameter**	**Min.**	**Max.**	**No. of values**
LASSO, Ridge (`glmnet` [52])	`lambda`	10^{-2}	10^{10}	100
GPLS (`caret` [107])	`ncomp`	1	5	5
SVM (`e1071` [39])	`cost`	0.01	3	6
	`gamma`	0	3	4
	`kernel`	–	–	4*
NNET (`nnet` [181])	`size`	1	13	7
	`decay`	10^{-4}	1	6
WKNN (`kknn` [77])	`k`	1	77	20
NB (`e1071` [39])	`laplace`	1	5	5

Algorithm (R package)	Parameter	Min.	Max.	No. of values
CART (`rpart` [174])	`cp`	0.001	0.1	5
C5.0 (`C50` [109])	`CF`	0	0.35	7
	`winnow`	`FALSE`	`TRUE`	2
	`rules`	`FALSE`	`TRUE`	2
RF (`ranger` [197])	`mtry`	4	100	7
	`min.node.size`	1	25	6
GBT (`xgboost` [30])	`eta`	0.01	0.4	4
	`max_depth`	1	3	3
	`colsample_bytree`	0.2	1	5
	`min_child_weight`	0.5	2	3
	`subsample`	0.2	1	3
	`nrounds`	50	250	3

8.2.5 Iterative Feature Elimination

We developed a feature selection wrapper that successively eliminates a subset of predictors that do not *positively* contribute to the performance of a model. The contribution of a predictor is computed using *model reliance* [48], which is a generalization of random forest permutation feature importance [21]. Model reliance estimates the merit of a predictor f toward a model ζ by comparing the classification error of ζ on the original training set $\mathbf{X}_{orig}$ with the classification error of ζ on a modified version of the training set $\mathbf{X}_{perm}$ where the values of f are randomly permuted. In particular, the model reliance MR of a model ζ on a predictor $f \in F$ is calculated as

$$MR(f, \zeta) = \frac{CE(y, \zeta(\mathbf{X}_{perm}))}{CE(y, \zeta(\mathbf{X}_{orig}))} \tag{8.6}$$

where CE is the classification error function that takes the true class labels y and a vector of predicted class labels, and returns the fraction of incorrectly classified observations. A high MR score represents a high dependence of the model on f, since shuffling the values of f increases the classification error. Conversely, a MR score smaller than 1 suggests that f is potentially adversarial to model performance, and its removal could increase model performance. Thus, our feature elimination wrapper starts by training a model on the full set of predictors, followed by an iterative step where the subset of adversarial predictors according to model reliance is removed, and a new model on the remaining predictors is trained. In the first iteration $i = 1$, an initial model ζ_1 is calculated on the full set of predictors $F_1 = F$. For each predictor $f \in F_i$, the model reliance $MR(f, \zeta_i)$ is calculated. Predictors with $f \in F_i : MR(f, \zeta_i) > 1$ are kept for iteration $i + 1$ while the remaining predictors are removed. This procedure is repeated either until all MR are smaller or equal to 1, i.e., $\forall f \in F_i : MR(f, \zeta_i) \leq 1$, or $F_{i+1} = F_i$. As random feature permutation introduces some statistical variability, we compute mean MR over 10 runs to obtain a more stable estimate.

8.2.6 Post-Hoc Interpretation

SHAP. To facilitate model interpretation, the model-agnostic post-hoc framework SHAP [125, 126] was used to assess feature importance for the CHA data. Briefly, the SHAP value $\phi_f(\zeta, x)$ expresses the estimated importance of a feature f to the prediction of model ζ for an instance x as change in the expected value of the prediction if for f the feature vector of x is observed

instead of being random. The SHAP framework composes the model prediction as sum of SHAP values of each feature, i.e., $\zeta(x) = \phi_0(\zeta, x) + \sum_{i=1}^{M} \phi_i(\zeta, x)$, where $\phi_0(\zeta, x)$ is the expected value of the model (bias) and M is the number of features.

SHAP values were calculated for the best model ζ_{opt} according to AUC. A ranking of T0 feature attribution towards ζ_{opt} was determined by calculating the average SHAP value magnitude over all instances, i.e., $A(j) = \sum_{i=1}^{N} |\phi_j(\zeta_{opt}, x)|$, where $A(j)$ is the attribution of the j-th feature. The $N \times M$ SHAP matrix was clustered with agglomerative hierarchical clustering to identify subgroups of patients with similar SHAP values.

PDP feature importance. We derive a *global* feature importance from the PD of a predictor. Our assumption is that predictors with high PD variability are more important. For example, consider the two PD curves in Figure 8.1 (c): the predicted response changes considerably with different predictor values for blue PD curve whereas the green PD curve is basically a flat line. Therefore, the predictor with the blue PD curve should have a higher importance score than the predictor with the green PD curve. We define partial dependence importance I of a predictor f as average of the magnitude of differences between consecutive values along the distribution of f, i.e.,

$$I_f = \frac{1}{k-1} \sum_{i}^{k-1} |PD(Q = s_i) - PD(Q = s_{i+1})| \tag{8.7}$$

where k is the number of (sampled) values from the distribution of f.

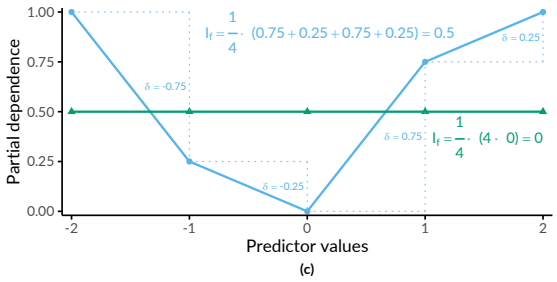

Figure 8.5: **Illustration of partial dependence importance.** Partial dependence importance I_f for 2 toy PD curves.

8.3 Results

In this section, we report our results on all three classification tasks, i.e., regarding CHA-Tinnitus (Section 8.3.1), CHA-Depression (Section 8.3.2) and AneurD (Section 8.3.3).

8.3.1 Results for CHA-Tinnitus

Correlational analysis. Figure 8.6 (a) shows all pairwise correlations among the predictors in T0. We identified two major subgroups with moderate to high intra-group correlations and low or negative inter-group correlations. The larger group (cf. upper black square in Figure 8.6 (a)) comprises 114 predictors (ca. 55.6%) representing negatively worded items and scores where higher values represent a higher disease burden, e.g. the ADSL_depression

and BI_overallcomplaints. Consequently, the smaller group (cf. lower black square in Figure 8.6 (a)) contains 47 predictors (ca. 22.9%) with positive wording, e.g. the SF8 mental health score (SF8_mental) and the BSF elevated mood score (BSF_mood). Predictors of one of the two subgroups exhibit a moderate to high negative correlation with the predictors of the other subgroup. Figure 8.6 (b) compares the correlation of the predictors with TQ_distress before (x-axis) and and after treatment (y-axis). Overall, only low to moderate bivariate correlations were observed, as all values are between -0.6 and +0.6. The average change in absolute correlation between T0 and T1 is 0.031. The change in absolute correlation is smaller than 0.067 for ca. 95% of the predictors (compare distance of the points to the diagonal line in Figure 8.6 (b)). For 137 out of 205 predictors (66.8%), absolute correlation decreased from T0 to T1. Median target-correlation of the questionnaires ADSL, BSF and BI (SF8) are greater (smaller) than +0.3 (-0.3) at both moments, respectively, and thus greater than for the remaining questionnaires. Figures 8.6 (c) and (d) reveal that predictors from ADSL, BSF, BI, SF8, TINSKAL and PSQ are among the top-20 predictors ranked by absolute correlation with TQ_distress in T0 and T1. The general depression score ADSL_depression shows largest correlation magnitude before ($\rho = 0.630$) and after treatment ($\rho = 0.564$). Figure 8.6 (e) shows the 10 predictors with the largest differences in correlation magnitudes between T0 and T1. Correlation before treatment is larger for each of these predictors.

Predictive performance of classification models. The performances of all 11 classifiers across each feature elimination iterations are shown in Figure 8.7. The gradient boosted trees model (GBT) yields highest AUC (iteration i = 7, AUC = 0.890 ± 0.04; mean±SD), using only 26 predictors (ca. 13%). The RIDGE classifier achievs second-best performance (i=2, AUC: 0.876 ± 0.05), relying on 127 features, followed by the random forest model (i=3, AUC: 0.872 ± 0.05) using 77 features. Classification using the best model (GBT, i=7) based on a probability threshold of 0.5 resulted in an accuracy of 0.86, a true positive rate (sensitivity) of 0.72, a true negative rate (specificity) of 0.88, a precision of 0.48 and a negative predictive value of 0.95.

When trained using a smaller feature space, each classifier produce at least one model with similar or even improved performance compared to the respective model learned on the whole set of predictors. In fact, with the exception of WKNN, all classification methods benefit from feature elimination as they produce their best model on a predictor subset (cf. Figure 8.7). For GBT, the gain in AUC from 185 features to 26 features (i = 11) is 0.01. This model achieves both maximum AUC and a good trade-off between high predictive performance and low model complexity, and we thus decided to further investigate this model.

Feature importance. For the best model, the attributions of the 26 selected features are shown in Figure 8.8 (a). Among the 26 features are 6 scores, 12 single items, 4 demographic features (number of visited doctors, university-level education, lower secondary education, tinnitus duration) and 4 features measuring the average time spent completing an item. The TINSKAL tinnitus impairment score (TINSKAL_impairment) represents the predictor with highest model attribution as it exhibits the highest average absolute SHAP value (change in log odds) of 0.448. The ADSL depressivity score (ADSL_depression) and a single question from ADSL (ADSL_adsl11: *"During the past week my sleep was restless."*) are ranked second and third most important, respectively. Remarkably, from each of the 9 questionnaires at least 1 feature was selected. Figure 8.8 (b) shows the patient-individual SHAP values for each predictor where point color depicts predictor value magnitude. The high attribution of TINSKAL_impairment is highlighted by the wide spread in the SHAP value distribution. For this predictor, high values generally correspond to an increased predicted probability of tinnitus decompensation. However, this trend is non-linear, since small values (light green to yellow) are associated with a SHAP value just slightly smaller than or equal to 0. Moreover, there is a large spread in SHAP value between ca. 0.7 and 1.2 for patients with high TINSKAL_impairment values opposed to the somewhat more dense bulk points representing patients with SHAP values between ca. -0.7 and

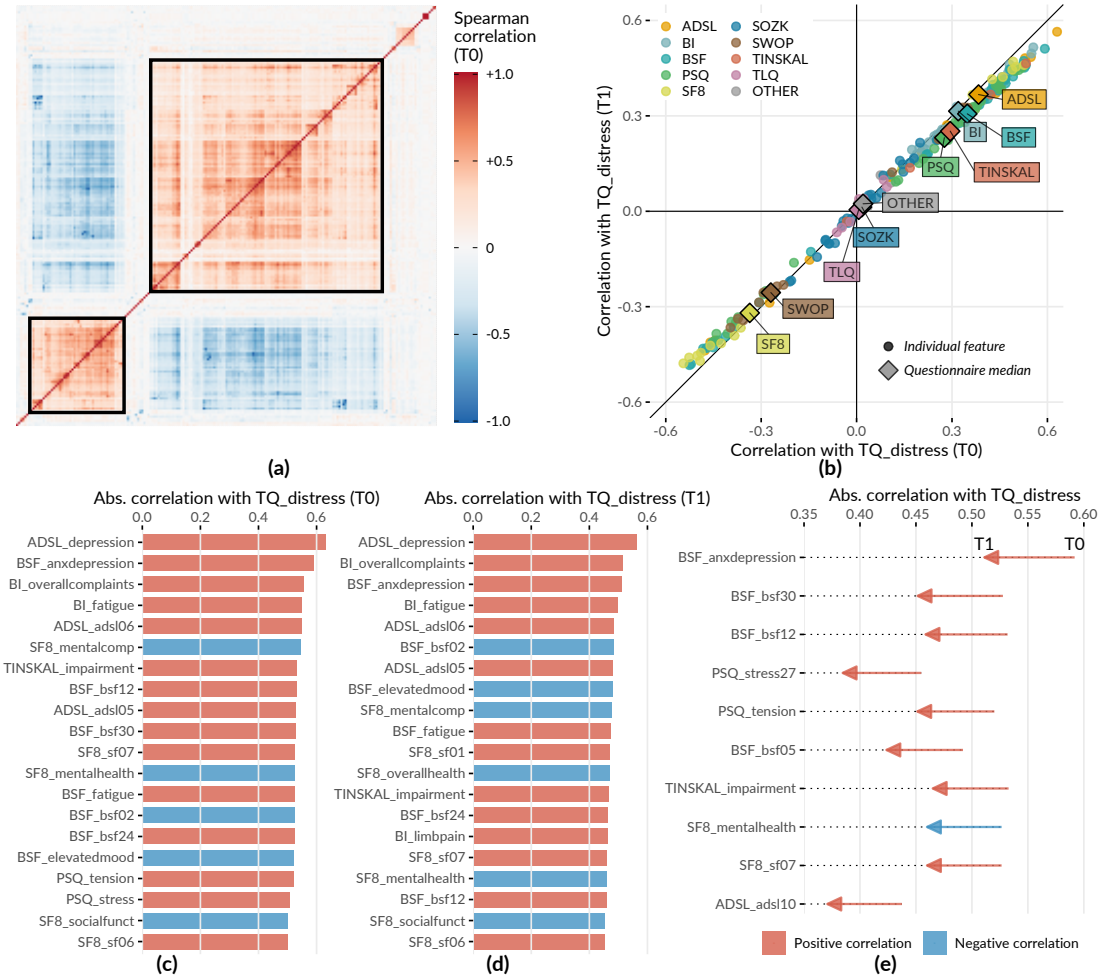

Figure 8.6: **Spearman correlation among predictors and correlation of predictors with TQ_distress in T0 and T1.** (a) The heatmap depicts the correlation coefficients for all pairs of predictors in T0. Predictors are arranged by the result of agglomerative hierarchical clustering with complete linkage. The two black squares depict two major subgroups of correlated predictors. (b) The relationship between each predictor with TQ_distress in T0 (x-axis) and in T1 (y-axis). The diamond symbol represents the median correlation of the predictors from the same questionnaire. (c) Top-20 predictors which exhibit highest absolute correlation with TQ_distress in T0. (d) Top-20 predictors which exhibit highest absolute correlation with TQ_distress in T1. (e) Top-10 predictors with the highest change in absolute correlation with TQ_distress from T0 to T1.

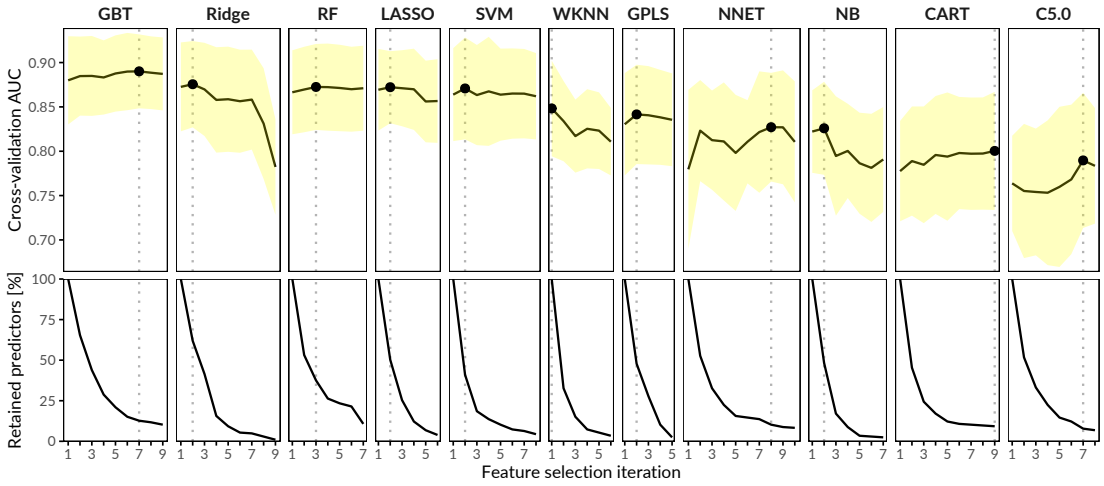

Figure 8.7: **Classification results for CHA-Tinnitus**. Average cross-validation AUC and relative number of retained predictors for each classifier with optimal hyperparameter configuration and for each feature selection iteration. Yellow ribbons depict standard deviation. Points highlight each classifier's run with maximum AUC. Classifiers are ordered by their maximum AUC from left to right.

-0.4. This could indicate that patients who report high tinnitus impairment are more difficult to classify. Further, it may suggest that visual analog scales are not robust enough to quantify tinnitus-related distress. This inference is supported by the SHAP feature dependence plot in Figure 8.9 (1) which juxtaposes the actual values of the predictor with the corresponding SHAP values for all patients and reveals a J-shaped relationship between them. More specifically, the predicted tinnitus-related distress is decreasing from 0 to 2.5, remains at a plateau from 2.5 to 4 and is increasing from 4 to its maximum value 10. Besides TINSKAL_impairment, the features ADSL_depression, TINSKAL_loudness, BI_overallcomplaints, BSF_timestamp and SWOP_pessimism also showed a non-linear relationship with respect to their SHAP values.

Even though several predictors exhibit only a low or moderate global importance, some of them have a high attribution towards model prediction for specific subgroups. For example, considering SOZK_lowersec, patients with lower secondary education have an average SHAP value of +0.5, whereas patient with different education levels have an average SHAP value of -0.1 and hence are more close to the population average (cf. Figure 8.8 (b), Figure 8.9 (13)). Most features show a monotonic relationship between actual values and SHAP values. For example, increasing values of the SF8 physical component score (SF8_physicalcomp) exhibit decreasing likelihood of predicted decompensated tinnitus with increasing physical health (cf. Figure 8.8 (b), Figure 8.9 (14)).

To investigate whether there are subgroups of patients which are similar with respect to how the model prediction can be explained by them, we clustered the patients based on the SHAP values. Figure 8.8 (c) shows stacked patient-individual SHAP values for the six predictors with highest average absolute SHAP values and the remaining predictors combined. According to Bayesian information criterion (cf. insetted plot in Figure 8.8 (c)), the optimal number of patients clusters with similar SHAP value patterns is 5. Clusters 1 and 5 comprise subgroups where the sum of SHAP values over all predictors is positive, see the horizontal lines in Figure 8.8 (c). Hence, these patients are more likely to be predicted with decompensated tinnitus. In comparison with the other subgroups, patients of clusters 1 and 5 reported higher degrees of tinnitus impairment, depressivity, anxiety, tinnitus loudness, sleeplessness, pessimism, psychosomatic complaints as

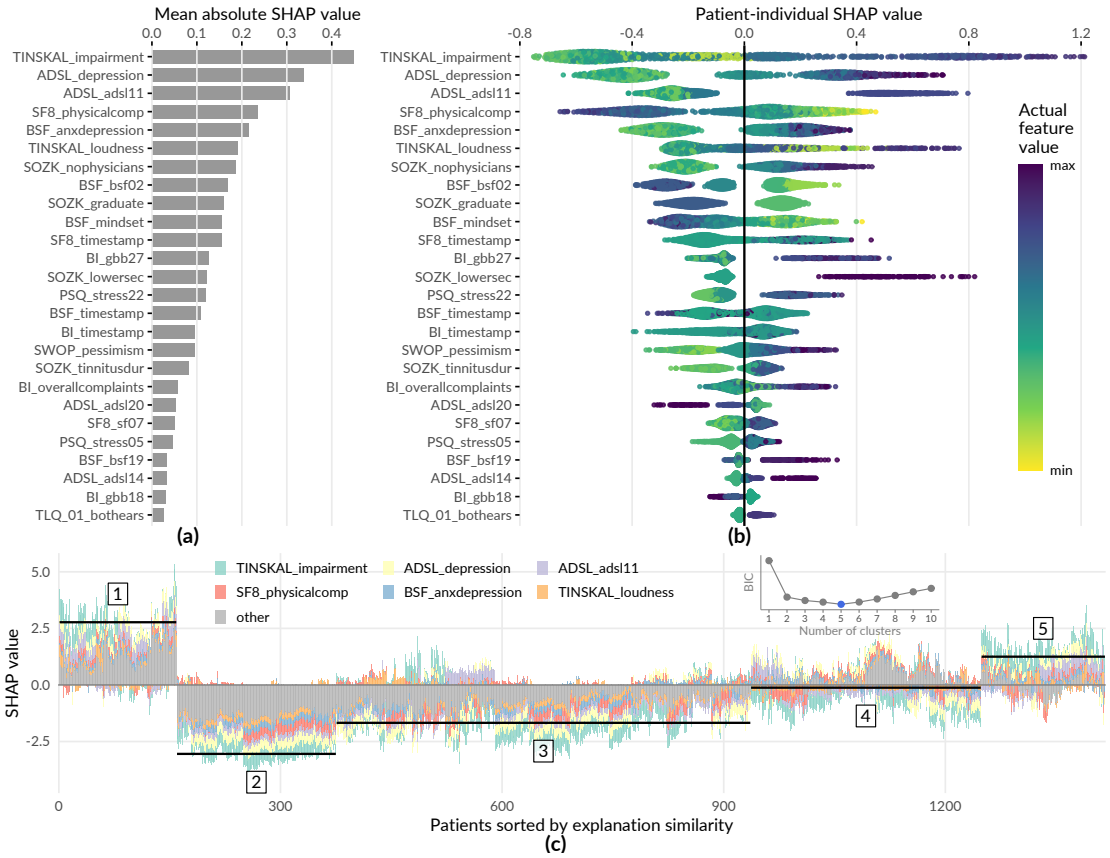

Figure 8.8: **SHAP analysis results for the best model (GBT, feature elimination iteration i = 7)**. (a) Global feature importance based on the mean absolute SHAP magnitude over all observations. Values depict absolute change in log odds where higher values indicate higher feature attribution towards the model. (b) Patient-individual SHAP values. Points represents the SHAP value of the predictor (y-axis) for an individual patient. The further afar a point from the vertical 0-baseline, the larger the attribution of the corresponding predictor value to the model prediction. Vertically offset points depict regions of high density (similar to a violin plot), i.e., there is a greater number of patients with similar SHAP values. Actual predictor values are mapped to point color. (c) Stacked patient-individual SHAP values for the 6 predictors with highest mean absolute SHAP values. Patients are ordered according to hierarchical clustering with Ward linkage. Black horizontal lines depict the average sum of SHAP values of the cluster members for k = 5 clusters. The inset plot shows that Bayesian information criterion is minimal for this number of clusters.

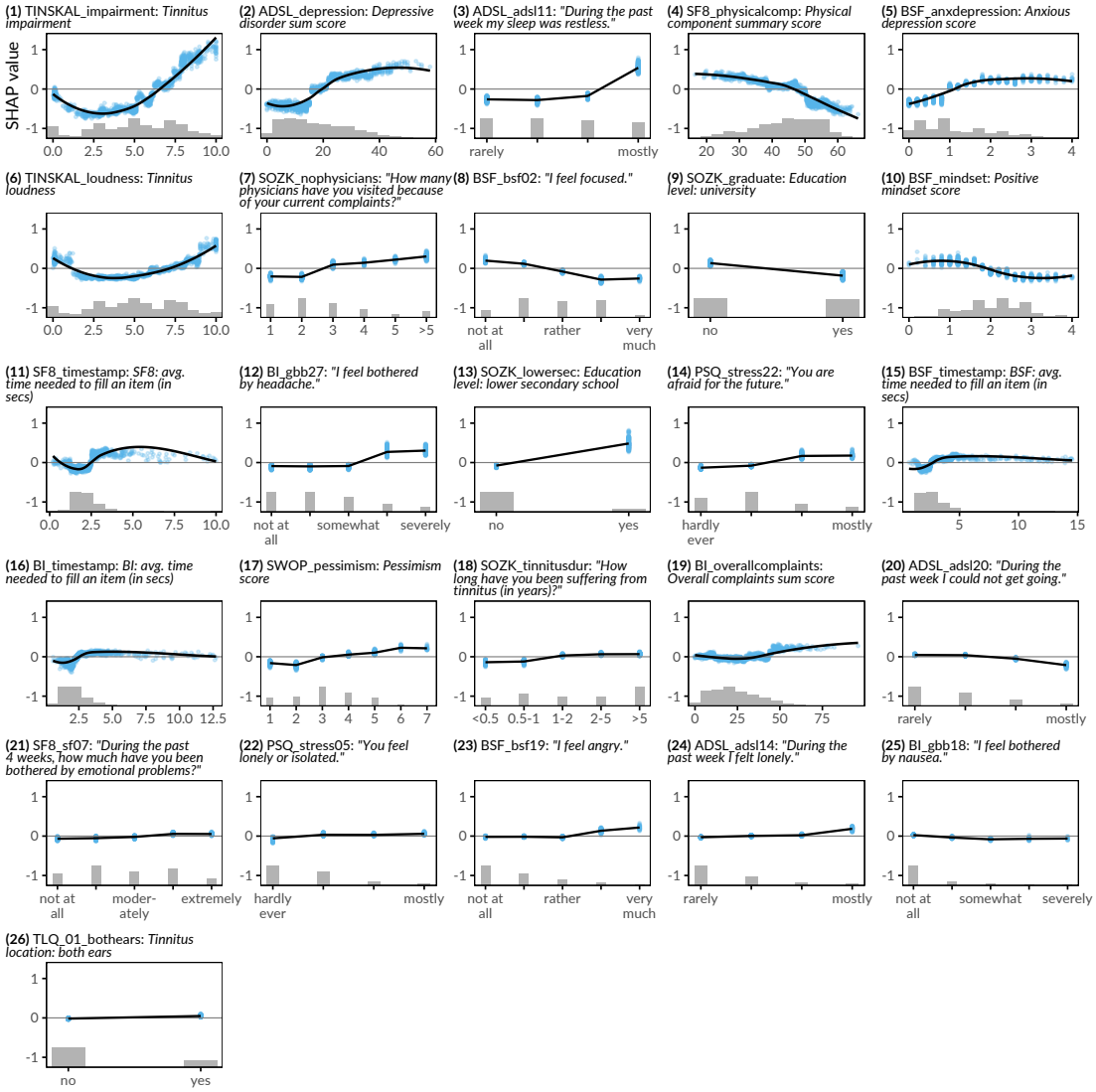

Figure 8.9: **SHAP feature dependence.** The relationship between the actual values of a predictor (x-axis) and corresponding SHAP values (y-axis) is shown as points representing a patient and as locally weighted scatterplot smoothing (LOWESS) curves indicating the overall trend. Predictors are ordered by mean absolute SHAP value (see Figure8.8 (a)). Gray histograms and bar charts depict the distributions of the predictors.

well as perceived levels of stress and social isolation. In general, patients of cluster 1 have slightly higher values across all predictors than patients of cluster 2. In addition, cluster 1 contains a higher fraction of patients with lower secondary education ("Hauptschule"), report more frequently occuring headaches, higher levels of fears for the future and a longer tinnitus duration. Cluster 3 is the largest subgroup comprising 39.6% of all patients. Together with cluster 2, these subgroups have the lowest predicted probability of tinnitus decomposition. Patients of cluster 2 and 3 report highest physical health and levels of determination. Cluster 4 is somewhat close to the prediction average where positive and negative SHAP values nearly even out. With respect to average patient-sum of SHAP values, cluster 3 lies in between cluster 2 and cluster 4.

8.3.2 Results for CHA-Depression

Predictive performance of classification models. Figure 8.10 depicts the performance of all classification methods across iterations. The LASSO classifier achieved maximum AUC over all classification algorithms (iteration i = 1, AUC = 0.867 ± 0.037; mean ± SD), followed by Ridge (i = 1, AUC = 0.864 ± 0.040) and GBT (i = 1, AUC = 0.862 ± 0.038). When considering only the best model per classifier, the models are similar in performance, ranging in AUC from 0.809 (C5.0) to 0.867 (LASSO).

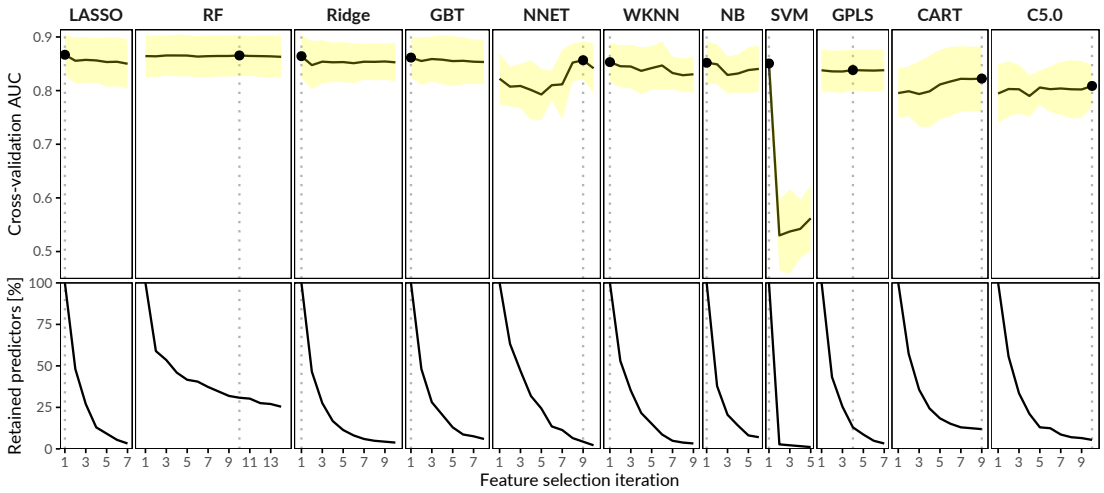

Figure 8.10: **Classification results for CHA-ADSL_depression.** Average cross-validation AUC and relative number of retained predictors for each classifier with optimal hyperparameter configuration and for each feature selection iteration. Yellow ribbons depict standard deviation. Points highlight each classifier's run with maximum AUC. Classifiers are ordered by their maximum AUC from left to right.

The best model (Lasso, i = 1) achieves an accuracy of 79%, a true positive rate (sensitivity) of 61%, a true negative rate (specificity) of 88%, a precision of 72% and a negative predictive value of 82% based on a probability threshold of 0.5. This final model includes 40 predictors with non-zero coefficients. Figure 8.11 shows the median model coefficient for these features across 10 cross-validation folds. From the ADSL questionnaire alone, 16 single items were included in the final model, including indicators of depressivity (ADSL_adsl09, ADSL_adsl18, ADSL_adsl12) perceived antipathy received from other people (ADSL_adsl19), sleeplessness (ADSL_adsl11), dejectedness (ADSL_adsl03), lack of appetite (ADSL_adsl02), confusion (ADSL_adsl05), anxiety (ADSL_adsl10, ADSL_adsl08), absence of self-respect (ADSL_adsl04, ADSL_adsl09), lack of vitality (ADSL_adsl09, ADSL_adsl09), taciturnity (ADSL_adsl13) and irritability

(ADSL_adsl01). Thus, this questionnaire contributed the highest number of predictors to the model. From the tinnitus-distress-oriented TQ, 5 predictors were selected. Further, the model used 5 predictors from the socio-demographics questionnaire (SOZK), including German nationality (SOZK_nationality) which has the highest absolute model coefficient, university level graduation (SOZK_graduate), tinnitus duration (SOZK_tinnitusdur), employment status (SOZK_job), marital status (SOZK_unmarried) and partnership status (SOZK_partnership).

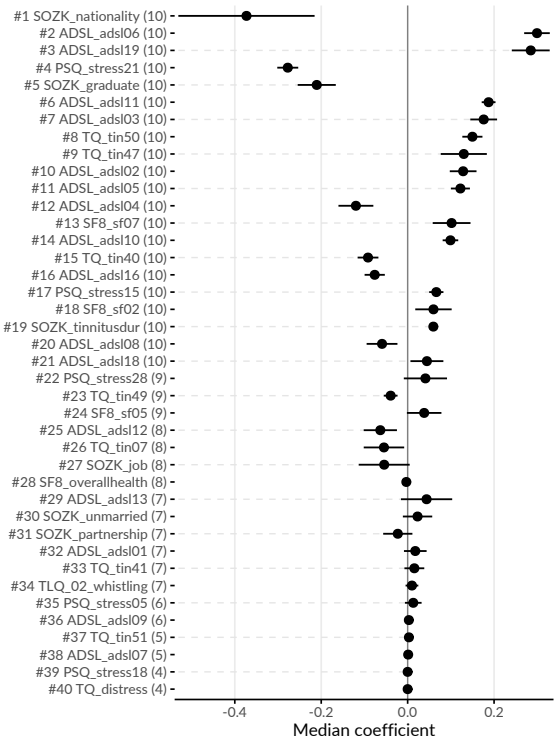

Figure 8.11: **Coefficients of LASSO model.** Cross-validation (CV) median (points) ± median absolute deviation (line ranges) of coefficients for the best LASSO model (i = 1). The frequency of non-zero coefficients in 10-fold CV is given in parentheses right to the predictor name. From 185 features in total, 40 features exhibit a non-zero model coefficient for at least one CV fold.

Table 8.3 provides a description for each of the predictors from Figure 8.11.

Effect of feature elimination of classification performance. All classifiers but SVM show high stability in performance on smaller feature subsets. From Figure 8.10, we see that for LASSO the difference in AUC when trained on 185 features (i = 1) vs. when trained on 6 features (i = 7) is only -0.017. Several classifiers benefited from feature selection in terms of predictive performance. For GPLS, NNET, CART, C5.0 and RF, max. AUC is achieved on a feature subset. Both decision tree variants CART and C5.0 gain most in performance from feature removal, since their respective max. AUC is obtained on the smallest predictor subset, with a cardinality of 22 and 10, respectively.

8.3.3 Results for AneurD

Figure 8.12 shows the classification results on each data subset. GBT achieves maximum AUC on *ALL* (cross-validation average 67.2% ± 1.8% standard deviation), followed by C5.0 (AUC 64.6% ± 1.9%) and GPLS (AUC 63.3% ± 1.2%). On the subset *SW*, SVM comes up best with

Table 8.3: **Most important features of LASSO model.** Predictors with highest absolute coefficient in the final LASSO model (iteration i = 1). From 185 predictors in total, these 40 predictors exhibit a non-zero model coefficient in at least one out of ten cross-validation folds.

Feature	Description	Coef.
SOZK_nationality	German nationality	-0.370
ADSL_adsl06	"During the past week I felt depressed."	0.309
ADSL_adsl19	"During the past week I felt that people disliked me."	0.288
PSQ_stress21	"You enjoy yourself."	-0.284
SOZK_graduate	Education level: university	-0.210
ADSL_adsl11	"During the past week my sleep was restless."	0.196
ADSL_adsl03	"During the past week I felt that I could not shake off the blues even with help from my family or friends."	0.175
TQ_tin50	"Because of the noises I am unable to enjoy the radio or television."	0.151
TQ_tin47	"I am a victim of my noises."	0.137
ADSL_adsl02	"During the past week I did not feel like eating; my appetite was poor."	0.132
ADSL_adsl05	"During the past week I had trouble keeping my mind on what I was doing."	0.132
SF8_sf07	"During the past 4 weeks, how much have you been bothered by emotional problems (...) ?"	0.125
ADSL_adsl10	"During the past week I felt fearful."	0.107
ADSL_adsl04	"During the past week I felt I was just as good as other people."	-0.107
TQ_tin40	"I am able to forget about the noises when I am doing something interesting."	-0.104
ADSL_adsl16	"During the past week I enjoyed life."	-0.085
PSQ_stress15	"Your problems seem to be piling up."	0.081
TQ_tin07	"Most of the time the noises are fairly quiet."	-0.069
ADSL_adsl08	"During the past week I felt hopeful about the future."	-0.064
SF8_sf02	"During the past 4 weeks, how much did physical health problems limit your physical activities (such as walking or climbing stairs)?"	0.059
SOZK_tinnitusdur	"How long have you been suffering from tinnitus (in years)?"	0.058
PSQ_stress28	"You feel loaded down with responsibility."	0.055
ADSL_adsl18	"During the past week I felt sad."	0.053
SOZK_job	Job status: currently employed	-0.050
ADSL_adsl13	"During the past week I talked less than usual."	0.049
TQ_tin49	"The noises are one of those problems in life you have to live with."	-0.048
ADSL_adsl12	"During the past week I was happy."	-0.046
SF8_sf05	"During the past 4 weeks, how much energy did you have?"	0.040
SOZK_unmarried	Unmarried	0.033
SOZK_partnership	In partnership	-0.032
TQ_tin41	"Because of the noises life seems to be getting on top of me."	0.031
PSQ_stress05	"You feel lonely or isolated."	0.025
ADSL_adsl01	"During the past week I was bothered by things that usually don't bother me."	0.017
TQ_tin51	"The noises sometimes produce a bad headache."	0.015
TLQ_02_whistling	Tinnitus noise: whistling	0.010
ADSL_adsl07	"During the past week I felt that everything I did was an effort."	0.010
ADSL_adsl09	"During the past week I thought my life had been a failure."	0.005
SF8_overallhealth	Overall health score	-0.003
PSQ_stress18	"You have many worries."	0.001
TQ_distress	Total tinnitus distress score	0.001

75.2% ± 5.7% AUC, slightly superior to GPLS (AUC 73.6% ± 4.4%) and NNET (AUC 71.6% ± 5.5%). For *BF*, WKNN yields the best (AUC 64.0% ± 1.1%) model, while GPLS (AUC 62.9% ± 2.6%) and RF (AUC 62.7% ± 2.3%) have similar yet slightly inferior generalization performances. Our results indicate that all classifiers yield better performance on the subset of sidewall aneurysms. Overall, none of the classification algorithms outperforms all others across all three subsets.

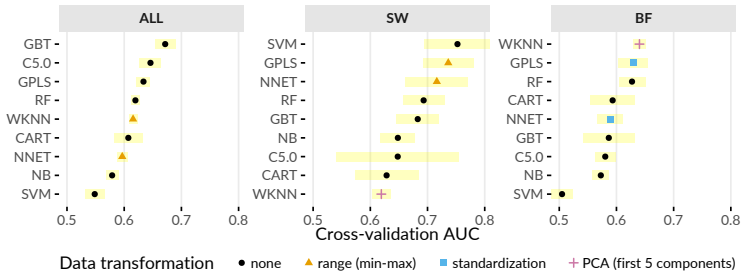

Figure 8.12: **Classification results for AneurD.** For each combination of data subset and classification algorithm, the performance of the run with the preprocessing transformation that achieves highest AUC is shown. SW = sidewall; BF = bifurcation.

With respect to PD importance, Figure 8.13 illustrates the high attribution of the angle parameter γ towards rupture status classification, as this feature is ranked first and third for the best models of *ALL* and *BF*. On the SVM model trained on the *SW* subset, ellipticity index (EI) is most important.

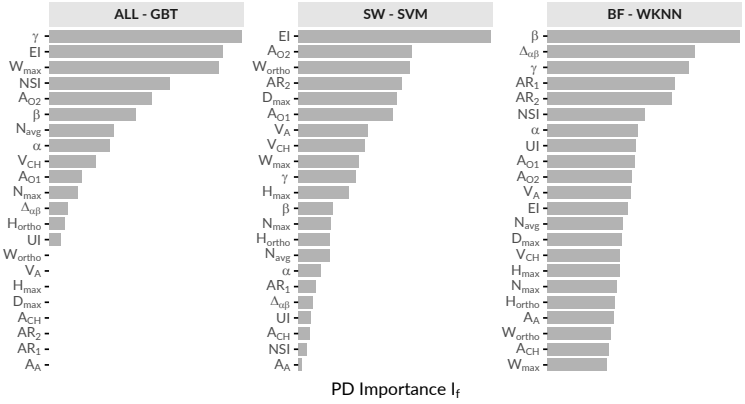

Figure 8.13: **Relative PD importance (AneurD).** PD importance for the best model of each data subset. Values are relative to the maximum PD importance. SW = sidewall; BF = bifurcation.

Figure 8.14 shows PDP and ICE curves for the most important predictors according to I_f for the best models on each data subset. All ICE curves of the GBT model and the SVM model (Figure 8.14 (a) and (b)) exhibit nearly identical trend but different intercept. In contrast, the ICE curves of WKNN (Figure8.14 (c)) appear more jittery. Apparently, the plots summarize major characteristics of the different model families. More specifically, the GBT model (Figure 8.14 (a)) produces jagged curves with distinct vertical cuts, representing the splits in the base decision trees of this tree ensemble. For example, the plot for γ shows 4 of such splits at {16.54, 49.26, 54.42, 64.35}. The SVM classifier (Figure 8.14 (b)) is a linear model, hence the ICE and PD curves are just lines with a fixed slope. The marginal model posterior of aneurysm rupture

increases with higher values of ellipticity index, max. width of the aneurysm body, aspect ratio H_{ortho}/N_{avg} and max. aneurysm diameter, whereas for the area of the ostium (variant 2), lower values are more indicative of a high rupture likelihood. For WKNN, the PDP is better able to clearly show the marginal posteriors of individual predictors than the ICE curves. This could be due to the property of WKNN being a "lazy" learner which does not produce an actual model but makes its predictions based on observation-individual similarity.

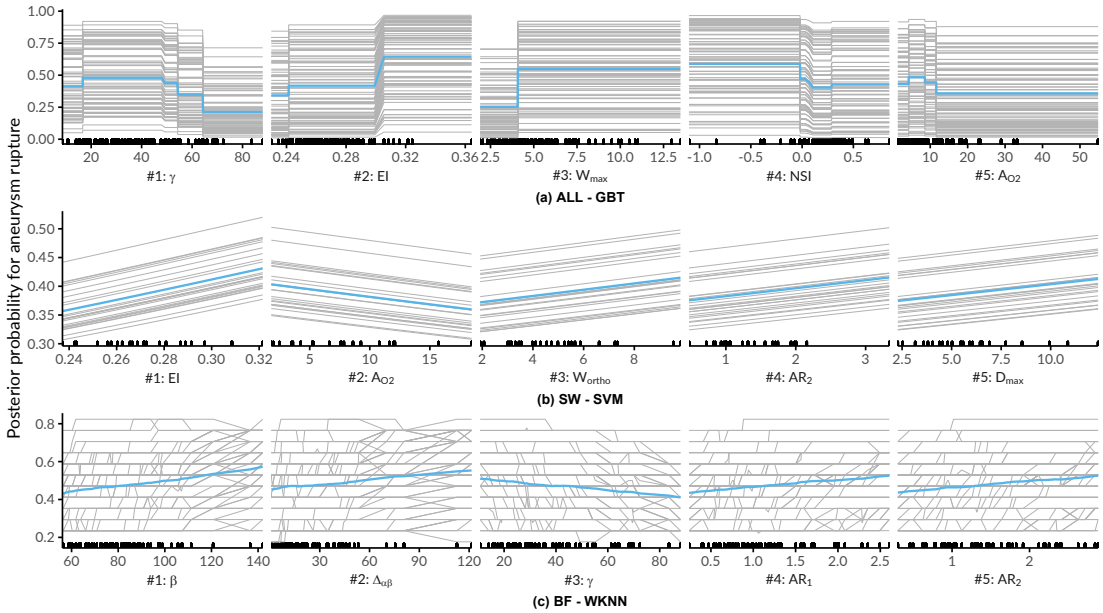

Figure 8.14: **Relative PD importance (AneurD).** PD importance for the best model of each data subset. Values are relative to the maximum PD importance. SW = sidewall; BF = bifurcation.

8.4 Discussion

In this section, we discuss our findings with respect to all three classification tasks, i.e., regarding CHA-Depression (Section 8.4.1), CHA-Tinnitus (Section 8.4.2) and AneurD (Section 8.4.3).

8.4.1 CHA-Depression

Machine learning has been used to build predictive models of depression severity based on structured patient interviews [124, 100]. We refrain from quantitative comparison with these studies due to differences in population characteristics and measurements. But how good is the best of our models actually? A reasonable baseline is a classifier that simply predicts depression status at time T0, which yields 79% accuracy. Our models outperform this baseline, although it is likely that they provide a good fit only for our sample, with patient subgroups from other centers yet to be studied. However, our models are a promising first step in supporting timely prediction of depression severity and selection of appropriate treatment with only a few questionnaire items.

Consistent with previous studies [114], we found a strong association between tinnitus distress and depression severity. Furthermore, predictors measuring perceived stress and demands were

found to be a significant contributor to depression in tinnitus patients [176]. The fact that predictors were selected from different questionnaires confirms the multifactoriality of depression, whose assessment requires the inclusion of different measurements. Therefore, concomitant emotional symptoms and other comorbidities must be taken into account to meet patient-specific needs. In a previous study [193], high sensitivity in detecting depression was achieved by using only a two-item questionnaire. One of the two items was *"During the past month, have you often been bothered by feeling down, depressed, or hopeless."*;[193] which is similar to the ADSL_adsl06 (*"In the past week, I have felt depressed."*), which has the second largest absolute coefficient in our best LASSO model.

Generally, care must be taken when interpreting the model coefficients: for example, we identified a strong relationship between non-German citizenship and depression severity (cf. Figure 8.11 and Table 8.3). Although some studies have reported ethnic differences in depression [152, 191], the occurrence of this item tends to suggest higher perceived social stress in patients of predominantly Turkish origin, due to higher unemployment rates, larger families, poorer housing conditions, etc. in this demographic group. Because only 5.0% of the cohort population were non-German citizens, these results could also be an effect of overfitting. Since the associated predictor in the first iteration of feature elimination has a model reliance score of less than 1.0, it is consequently omitted from the sparser models.

Regarding the stability of the models on smaller feature sets, our results show that simpler models are only slightly inferior to the most predictive model. More specifically, most classification methods show an improvement in AUC as the number of predictors decreases. In fact, 5 of 11 classifiers improved even by feature selection, i.e., the AUC in the second or later iteration was superior to the AUC in the first iteration (where all 205 predictors are used). For example, the two decision tree variants achieved the highest performance on the smallest feature subset in each case. Regarding the LASSO classifier, which showed the best performance, it is encouraging that only 6 predictors from 4 questionnaires showed similar performance (AUC = 0.850) compared to the best overall model (AUC = 0.867). It is noteworthy that neither predictors of tinnitus localization and quality nor sociodemographic predictors were included in this model. This finding could be used to reduce the number of questions or entire questionnaires that patients must answer before and after treatment. psychological or physical stress for a subject undergoing an examination (e.g., a painful biopsy vs. a blood test). For example, Yu et al. [199] perform feature selection under a budget, where the cost of feature acquisition is derived from suggestions by medical experts based on total financial burden, patient privacy, and patient inconvenience. Kachuee et al. [97] derive feature costs based on the convenience of answering questions, performing medical exams, and blood and urine tests.

In terms of clinical relevance, our results should be a first step to guide clinicians in making treatment decisions regarding clinical depression in patients with chronic tinnitus. The models could be used to design an appropriate treatment pathway. However, before using the models in practice, one must be aware that they are trained on cross-sectional data, i.e., the models separate subclinical and clinical depression based on questionnaire responses and sociodemographic data before treatment. Also, one must keep in mind that the treatment was a 7-day treatment and the response is depression status *after treatment*.

There are also some limitations to our approach. First, our models might be subject to selection bias because patients who did not complete all seven questionnaires both at admission *and* after treatment were excluded from our analyses. However, we do not consider these data as "missing values" because this could lead to the problematic suggestion of using imputation methods. We cannot use imputation because (i) a proportion of patients did not complete the entire questionnaire (rather than individual items) and (ii) we do not know whether the data are missing at random. However, because the number of patients is large, we believe our results are sufficiently robust. In future work, we will investigate possible systematic differences between

included and excluded patients. The exclusion of patients who dropped out of completing the questionnaires prematurely, partly because of a gradual loss of motivation, technical unfamiliarity with the computer, or possible interruptions by staff to complete other baseline assessments, could lead to selection bias. Because the patient population was from only one hospital, future work involves external validation of the models on data from different populations and hospitals. Because the use of cross-sectional data limits the interpretation of the prediction of depression severity beyond the end of therapy, future work will need to validate the models with longitudinal data.

Another potential limitation is the greedy process of our iterative feature selection wrapper, which can miss global optima as a result. At each iteration, predictors that prevent the model from classifying correctly are removed from the feature set. Once a predictor is eliminated, it cannot be included in any subsequent iteration. However, it is possible that including a predictor that was removed in an early iteration could lead to a better model in a later iteration. A possible solution would be a mechanism to backtrack or revisit earlier iterations if it turns out that some of the removed predictors actually contributed positively to model performance. Alternatively, the MR cutoff value for discarding features (which was set to 1 in our experiments) could be subjected to hyperparameter tuning. Future work therefore includes comparison with other feature selection algorithms.

8.4.2 CHA-Tinnitus

We trained classification models to predict tinnitus-related distress after multimodal treatment (T1) in patients with chronic tinnitus based on self-report questionnaires data acquired before treatment (T0). The gradient boosted trees model which uses 26 (12.7%) from a total of 205 predictors separates patients with “compensated” vs. “decompensated” tinnitus with best AUC.

Among these features are measurements that describe a variety of psychological and psychosomatic patient characteristics as well as socio-demographics and therefore confirming the multifactorial nature of tinnitus-related distress. These characteristics were used for the phenotyping in Chapter 5. Additionally, the predictors can be investigated in a followup studies of how such characteristics influence treatment success. As expected, predictors that are directly linked to tinnitus quality show high model attribution, such as the degree of perceived tinnitus impairment and loudness. At the same time, depression, attitudinal factors (self-efficacy, pessimism, complaint tendency), sleep problems, educational level, tinnitus location and duration emerged as highly important for the model prediction as well.

Quantitative predictors, such as tinnitus impairment and loudness, show non-monotonic relationships with respect to the predicted outcome. Notably, very low self-reported impairment or loudness measured by visual analogue scales do not generally indicate low tinnitus-related distress measured by the TQ. One explanation is that simple measurements like TINSKAL_impairment and TINSKAL_loudness are less robust and show higher variability than a compound scale that combines multiple single questionnaire items. These findings could be investigated further, e.g., whether there is a relationship towards a subgroup of patients that were more fatigued and thus not less thoroughly filling a large number of questionnaires.

Our results confirm the intricate interplay between depression and tinnitus-related distress that was elucidated by numerous previous studies [41, 50, 69, 115, 155]. For our best model, an ADSL score of more than 20 is associated with an increased predicted risk of tinnitus decompensation (cf. Figure 8.9 (2)) which is close to the cutoff of clinical relevance of depression [76].

In the context of parsimonious learning, a general strategy is to determine the set of predictors which is as small as possible and where the inclusion of any other predictor does not yield in a considerable improvement in performance. So how many predictors are really necessary for an

accurate tinnitus distress prediction? Figure 8.15 (a) illustrates the change in performance for a GBT classifier when predictors are iteratively added to the feature space in the order of their SHAP values with respect to our best model. A model that uses only the TINSKAL_impairment achieves AUC = 0.79 ± 0.06. Adding ADSL_depression leads to an improvement in AUC of 0.06. However, none of the remaining 24 predictors results in an improvement of more than 0.01, respectively. Moreover, only 3 predictors are necessary for a model with AUC = 0.85, 8 predictors for a model with AUC = 0.87 and 15 predictors for a model with AUC = 0.89 (cf. Figure 8.15 (a)).

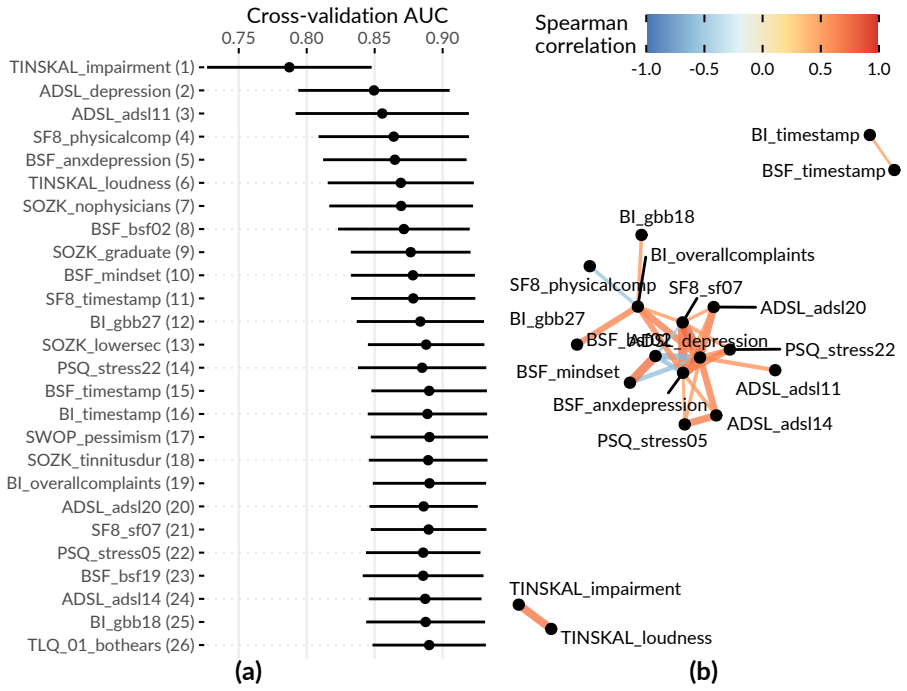

Figure 8.15: **Cumulative feature contribution & correlation network.** (a) Cross-validation AUC (average ± standard deviation) of a GBT model trained on the feature subset comprising the predictors denoted on the y-axis up to that iteration. The ordering of features is according to mean absolute SHAP value (cf. Figure 8.8~(a)). (b) Network illustrating 3 groups of features among the 26 selected predictors of the best model with high intra-group correlation ($|\rho| \geq 0.5$). 8 predictors (predominantly from SOZK) without any moderate to high pairwise correlation are not shown.

One potential explanation could be multicollinearity among groups of predictors. Figure 8.15 (b) shows a network of 3 predictor groups among the 26 features of the best model. For example, the features TINSKAL_impairment and TINSKAL_loudness are moderately correlated (Spearman correlation $\rho = 0.69$), which raises the question of whether one of the two predictors could be removed without a considerable loss in AUC. The largest subgroup spanning 14 features involves descriptors of depression, perceived stress and reported physical health. In future work, an investigation of possible interaction effects among these moderately to strongly correlated features could be investigated, to better understand why all of them were selected and to determine whether some of them could be removed to achieve a better trade-off between model accuracy and complexity.

Our workflow leverages the potential of machine learning for identifying key predictors from a variety of features collected before treatment for post-treatment tinnitus compensation, by

ensuring that every potential predictor is included in the analysis, and by the internal validation of the classification models using cross-validation and hyperparameter tuning. Furthermore, by selecting a variety of classification algorithm families, both linear and nonlinear relationships between a feature and outcome could be identified. A limitation of this hypothesis-free approach is that the learned models could contain features that quantify the same or similar patient characteristics. For example, the best model in this study included the two highly correlated features ADSL_depression and BSF_anx_depression (anxious depressiveness score). While the inclusion of both features contributed to model performance, from a medical perspective, a predictive model with only certain features might be more beneficial. Preselecting features to avoid multicollinearity could be a direction for future work.

Finally, the exclusion of 2,701 out of 4,117 patients (65.6%) who did not complete *all* 10 questionnaires could have resulted in selection bias. Many patients spent more than one hour completing the questionnaire on a dedicated minicomputer and were therefore more likely to drop out of the completion process. Completers were slightly younger than non-completers (mean age 49.8 $\pm$ 12.2 vs. 51.7 $\pm$ 13.8), were more likely to have the highest German school degree "Abitur" (48.2% vs. 42.0%) and had been suffering from tinnitus longer ($>$ 5 years: 33.3% vs. 25.1%). In future work we intend to investigate to what extent insights from completers can be used on subsamples of non-completers. Therefore, we can use the DIVA framework of Hielscher et al. [81]. However, psychological treatment approaches are likely to benefit only those who report psychological problems prior to tinnitus perception or in association with tinnitus perception.

8.4.3 AneurD

Our classification results are promising, as morphological parameters alone can provide models with moderate power. Because previous studies have found that hemodynamic parameters are also predictive [26, 13], future work includes exploring the potential of combining morphologic and hemodynamic features for classification of rupture status. In addition, because our focus for now has been on quantifying the merit of morphologic parameters, we have ignored demographic characteristics, such as age and sex, which also correlate strongly with aneurysm rupture [36]. We expect that adding these patient characteristics will further improve classification performance.

PDP analysis showed that for the best model (gradient boosted trees), the parameters angle at the dome point γ, ellipticity index EI, maximum aneurysm width W_{max}, nonsphericity index NSI, and aneurysm area A_{O2} had the highest attribution (see Figure 8.13 (a)). These differed from those found for two subsets of sidewall and bifurcation aneurysms, respectively. Figure 8.14 shows that none of the features appear among the top 5 predictors for sidewall aneurysms, bifurcation aneurysms, and the overall data set. This is also partly due to the fact that the family of the best model is different for each of the subsets. Consequently, Figure 8.16 shows that the PDP curves for the top 5 features on ALL differ substantially. Therefore, we argue that PDPs are more appropriate for *intra*-model comparisons of feature attributions.

We observed that classification performance is consistently higher for the subset of sidewall aneurysms vs. bifurcation aneurysms, and that different parameters were found to have high model attribution. This could be partially due to the rather small sample size or the already mentioned differences in model families. However, Baharoglu et al. [12] identified significant differences between sidewall and bifurcation aneurysms with respect to morphological parameters, and how these parameters can predict rupture status.

Some of our findings also suggest some form of higher-level interactions between groups of features. For example, the ellipticity index (EI)was found second most important for ALL -

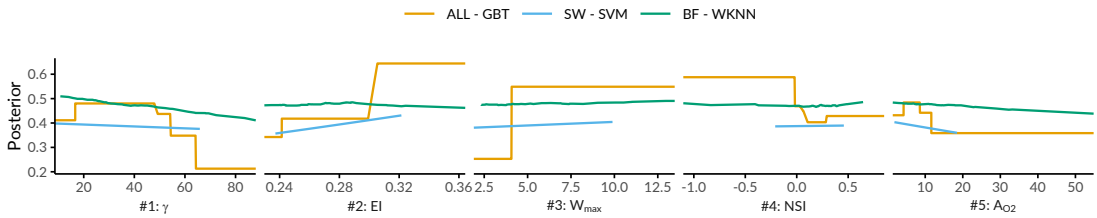

Figure 8.16: **PDP curves for top-5 predictors on ALL - GBT for the best models on each data subset.**

GBT and most important for SW - SVM, although differences in EI between unruptured and ruptured aneurysms are not significant ($p = 0.323$, Wilcoxon rank sum test, $\alpha = 0.01$).

There are some limitations to our analysis. The small sample size, especially for the subset of sidewall aneurysms (N=24), could lead to overfitting. In future work, we would like to retrain our models on a larger number of datasets, and incorporate a wider variety of predictors, such as hemodynamic and demographic features, as mentioned above. A further limitation concerns the validity of the class label. Samples that were labeled as unruptured could have ruptured at a later moment. Further, we would like to investigate samples with high classification error in more detail. Here, our goal is to derive descriptions of aneurysms subgroups which are hard to classify, in order to better understand reasons for misclassification, and to signalize to the medical expert that a manual diagnosis is necessary.

8.5 Conclusion

In medical applications, black-box models are becoming increasingly popular due to their high predictive power. However, due to their opacity, a post-modeling step is required to extract actionable insights from them.
We presented a machine learning workflow for classification and post-hoc interpretation alongside dataset-specific steps for three medical applications. While the variety of classification algorithms to be created is identical for all datasets, we chose the interpretation method based on the opacity of the model family that achieves maximum performance, and on dimensionality. For CHA-tinnitus, gradient-boosted trees performed best, and we used Shapely value explanations to obtain feature importance values at model, subpopulation, and observation levels. For CHA depression, LASSO was found to achieve the best generalization performance, so we used the intrinsically interpretable model coefficients. For AneurD, we used PD importance instead of SHAP values, although gradient-boosted trees provide the best model because the number of observations is too small to produce reliable importance scores at the subpopulation or observation level.

TODO: some discussion on the robustness

Chapter 9

Subpopulation-Specific Learning and Post-Hoc Model Interpretation

Brief Chapter Summary

We present a workflow to examine how subpopulations differ with respect to their most predictive characteristics in temporal data. To do so, we derive a post-hoc interpretation measure to assess the difference in association of predictors between two subpopulations. We report results for CHA on gender differences (subpopulations of female and male patients) for the two outcomes of tinnitus-related distress and depression, and the effect of treatment for both outcomes.

This chapter is partly based on:
Uli Niemann, Benjamin Boecking, Petra Brueggemann, Birgit Mazurek, and Myra Spiliopoulou. "Gender-Specific Differences in Patients With Chronic Tinnitus – Baseline Characteristics and Treatment Effects". In: *Frontiers in Neuroscience* 14 (2020), p. 487. DOI: 10.3389/fnins.2020.00487.

9.1 Motivation and Comparison to Related Work

to be written

9.2 Comparing Differences in Feature Importance between Two Subpopulations

to be written

9.3 Workflow

9.3.1 Learning Tasks

to be written

9.3.2 Model Evaluation and Hyperparameter Tuning

9.4 Results

to be written

9.5 Conclusions on Subpopulation-Specific Differences in Feature Importance

to be written

Part IV

SUMMARY

Chapter 10

Conclusion and Future Work

to be written

10.1 Research Results for Medical Expert-Guided Knowledge Discovery

to be written

10.2 Future Work

to be written

Abbreviations

AI	Artificial Intelligence
DM	Data Mining
ML	Machine Learning

Appendix A

Variables selected for phenotyping

- ACSA_qualityoflife*: Quality of life during the last 2 weeks
- ADSL_depression: Depressive disorder sum score
- BI_abdominalsymptoms: Abdominal symptoms score
- BI_fatigue: Fatigue score
- BI_heartsymptoms: Heart symptoms score
- BI_limbpain: Limb pain score
- BI_overallcomplaints: Overall complaints sum score
- BSF_anger: Anger score
- BSF_anxdepression: Anxious depression score
- BSF_apathy: Apathy score
- BSF_elevatedmood*: Elevated mood score
- BSF_fatigue: Fatigue score
- BSF_mindset*: Positive mindset score
- ISR_additionalitems: Additional items score
- ISR_anxiety: Anxiety score
- ISR_compulsivesyn: Obsessive-compulsive syndrome score
- ISR_depression: Depression score
- ISR_eatingdisorder: Eating disorder score
- ISR_somatosyn: Somatoform syndrome score
- ISR_totalpsychiatricsyn: Total psychiatric syndrome score
- PHQK_depression: Presence of depression
- PHQK_panicsyn: Presence of panic syndrome
- PSQ_demand: Demand score
- PSQ_joy*: Joy score
- PSQ_stress: Total perceived stress sum score
- PSQ_tension: Tension score
- PSQ_worries: Worries score
- SES_affectivepain: Affective pain
- SES_sensoricpain: Sensoric pain
- SF8_bodilyhealth*: Bodily health score
- SF8_mentalcomp*: Mental component summary score
- SF8_mentalhealth*: Mental health score
- SF8_overallhealth*: Overall health score

- SF8_physicalcomp*: Physical component summary score
- SF8_physicalfunct*: Physical functioning score
- SF8_roleemotional*: Role emotional score
- SF8_rolephysical*: Role physical score
- SF8_socialfunct*: Social functioning score
- SF8_vitality*: Vitality score
- SSKAL_painfrequency: Visual analog scale pain frequency
- SSKAL_painimpairment: Visual analog scale pain impairment
- SSKAL_painseverity: Visual analog scale pain severity
- SWOP_optimism*: Optimism score
- SWOP_pessimism: Pessimism score
- SWOP_selfefficacy*: Self-efficacy score
- TINSKAL_frequency: Tinnitus frequency
- TINSKAL_impairment: Tinnitus impairment
- TINSKAL_loudness: Tinnitus loudness
- TLQ_01_bothears: Tinnitus location: both ears
- TLQ_01_entirehead: Tinnitus location: entire head
- TLQ_01_leftear: Tinnitus location: left ear
- TLQ_01_rightear: Tinnitus location: right ear
- TLQ_02_hissing: Tinnitus noise: hissing
- TLQ_02_ringing: Tinnitus noise: ringing
- TLQ_02_rustling: Tinnitus noise: rustling
- TLQ_02_whistling: Tinnitus noise: whistling
- TQ_auditoryperceptdiff: Auditory perceptual difficulties score
- TQ_cognitivedistress: Cognitive distress score
- TQ_distress: Total tinnitus distress score
- TQ_emodistress: Emotional distress score
- TQ_intrusiveness: Intrusiveness score
- TQ_psychodistress: Psychological distress score
- TQ_sleepdisturbances: Sleep disturbances score
- TQ_somacomplaints: Somatic complaints score

Bibliography

[1] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] Shiva Alemzadeh et al. "Subpopulation Discovery and Validation in Epidemiological Data". In: *EuroVis Workshop on Visual Analytics (EuroVA)*. Ed. by Michael Sedlmair and Christian Tominski. The Eurographics Association, 2017. ISBN: 978-3-03868-042-0. DOI: 10.2312/eurova.20171118. URL: https://doi.org/10.2312/eurova.20171118.

[3] David Alvarez-Melis and Tommi S Jaakkola. "On the robustness of interpretability methods". In: *arXiv preprint arXiv:1806.08049* (2018).

[4] Luca Antiga et al. "An image-based modeling framework for patient-specific computational hemodynamics". In: *MBEC* 46.11 (2008), p. 1097.

[5] Catiele Antunes, Mohammadreza Azadfard, and Mohit Gupta. "Fatty liver". In: *StatPearls Publishing* (2019).

[6] Martin Atzmüller. "Subgroup discovery". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.1 (2015), pp. 35–49.

[7] Martin Atzmüller and Florian Lemmerich. "VIKAMINE–open-source subgroup discovery, pattern mining, and analytics". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2012, pp. 842–845.

[8] Martin Atzmüller and Frank Puppe. "SD-Map–A fast algorithm for exhaustive subgroup discovery". In: *Proc. of PKDD*. Springer. 2006, pp. 6–17.

[9] Peter C Austin et al. "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes". In: *J. of Clinical Epidemiology* 66.4 (2013), pp. 398–407.

[10] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[11] David Baguley, Don McFerran, and Deborah Hall. "Tinnitus". In: *The Lancet* 382.9904 (2013), pp. 1600–1607.

[12] Merih I Baharoglu et al. "Identification of a dichotomy in morphological predictors of rupture status between sidewall-and bifurcation-type intracranial aneurysms". In: *Journal of neurosurgery* 116.4 (2012), pp. 871–881.

[13] Philipp Berg and Oliver Beuing. "Multiple intracranial aneurysms: a direct hemodynamic comparison between ruptured and unruptured vessel malformations". In: *International journal of computer assisted radiology and surgery* 13.1 (2018), pp. 83–93.

[14] Jürgen Bernard et al. "A visual-interactive system for prostate cancer cohort analysis". In: *IEEE computer graphics and applications* 35.3 (2015), pp. 44–55. DOI: 10.1109/MCG.2015.49.

[15] Jan L. Bernheim and Marc Buyse. "The Anamnestic Comparative Self-Assessment for Measuring the Subjective Quality of Life of Cancer Patients". In: *Journal of Psychosocial Oncology* 1.4 (1993), pp. 25–38. DOI: 10.1300/J077v01n04_03.

[16] Jay M. Bhatt, Neil Bhattacharyya, and Harrison W. Lin. "Relationships Between Tinnitus and the Prevalence of Anxiety and Depression". In: *The Laryngoscope* 127.2 (2017), pp. 466–469. DOI: 10.1002/lary.26107.

[17] Bernd Bischl et al. "mlr: Machine Learning in R". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5938–5942.

[18] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proc. of Workshop on Computational Learning Theory*. ACM. 1992, pp. 144–152.

[19] A. J. Boulton et al. "The global burden of diabetic foot disease". In: *Lancet* 366.9498 (2005), pp. 1719–24. ISSN: 1474-547X (Electronic) 0140-6736 (Linking). DOI: 10.1016/S0140-6736(05)67698-2. URL: https://www.ncbi.nlm.nih.gov/pubmed/16291066.

[20] L. Breiman et al. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[21] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[22] Petra Brüggemann et al. "Impact of multiple factors on the degree of tinnitus distress". In: *Frontiers in human neuroscience* 10 (2016), p. 341.

[23] M Bullinger and M Morfeld. "Der SF-36 Health Survey". In: *Gesundheitsökonomische Evaluationen*. Springer, 2008, pp. 387–402.

[24] Cristóbal J Carmona et al. "Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department". In: *Soft Computing* 15.12 (2011), pp. 2435–2448.

[25] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. "Machine Learning Interpretability: A Survey onMethods and Metrics". In: *Electronics* 8.8 (2019), p. 832. DOI: 10.3390/electronics8080832.

[26] J.R. Cebral et al. "Quantitative Characterization of the Hemodynamic Environment in Ruptured and Unruptured Brain Aneurysms". In: *American Journal of Neuroradiology* 32.1 (2011), pp. 145–151.

[27] Christopher R Cederroth et al. "Towards an understanding of tinnitus heterogeneity". In: *Frontiers in aging neuroscience* 11 (2019), p. 53.

[28] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

[29] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16.1 (2002), pp. 321–357.

[30] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[31] Peter Clark and Tim Niblett. "The CN2 induction algorithm". In: *Machine learning* 3.4 (1989), pp. 261–283.

[32] W. W. Cohen. "Fast effective rule induction". In: *International Conference on Machine Learning*. 1995, pp. 115–123.

[33] Carlo Combi, Elpida Keravnou-Papailiou, and Yuval Shahar. *Temporal Information Systems in Medicine*. Springer, 2010.

[34] Alberto Corvo et al. "Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology". In: *IEEE Transactions on Visualization and Computer Graphics* (2020). DOI: 10.1109/TVCG.2020.2990336.

[35] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Vol. 1. Cambridge University Press, 1997.

[36] Felicitas J Detmer et al. "Development and internal validation of an aneurysm rupture probability model based on patient characteristics and aneurysm location, morphology, and hemodynamics". In: *International Journal of Computer Assisted Radiology and Surgery* 13.11 (2018), pp. 1767–1779. DOI: 10.1007/s11548-018-1837-0.

[37] Sujan Dhar et al. "Morphology Parameters for Intracranial Aneurysm Rupture Risk Assessment". In: *Neurosurgery* 63.2 (2008), pp. 185–197.

[38] Lee R Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

[39] Evgenia Dimitriadou et al. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-25*. Vienna, Austria, 2011. URL: http://cran.r-project.org/package=e1071.

[40] Beiying Ding and Robert Gentleman. "Classification using generalized partial least squares". In: *Journal of Computational and Graphical Statistics* 14.2 (2005), pp. 280–298.

[41] Robert A Dobie. "Depression and tinnitus". In: *Otolaryngologic Clinics of North America* 36.2 (2003), pp. 383–388.

[42] Shahram Ebadollahi et al. "Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics". In: *AMIA annual symposium proceedings*. Vol. 2010. American Medical Informatics Association. 2010, p. 192.

[43] Alev Akyol Erikci et al. "The effect of subclinical hypothyroidism on platelet parameters". In: *Hematology* 14.2 (2009), pp. 115–117.

[44] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[45] A Shojaie Fard, M Esmaelzadeh, and B Larijani. "Assessment and treatment of diabetic foot ulcer". In: *International journal of clinical practice* 61.11 (2007), pp. 1931–1938.

[46] Usama M. Fayyad and Keki B Irani. "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning". In: *Proceedings of the 13th International Joint Converence on Artificial Intelligence*. Vol. 2. 1993, pp. 1022–1027.

[47] Serafino Fazio et al. "Effects of thyroid hormone on the cardiovascular system". In: *Recent progress in hormone research* 59.1 (2004), pp. 31–50.

[48] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance". In: *arXiv preprint arXiv:1801.01489* (2018).

[49] Herbert Fliege et al. "The Perceived Stress Questionnaire (PSQ) reconsidered: validation and reference values from different clinical and healthy adult samples". In: *Psychosomatic medicine* 67.1 (2005), pp. 78–88.

[50] Robert L Folmer et al. "Tinnitus severity, loudness, and depression". In: *Otolaryngology—Head and Neck Surgery* 121.1 (1999), pp. 48–51.

[51] Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 2016.

[52] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[53] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of Statistics* (2001), pp. 1189–1232.

[54] Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of rule learning*. Springer Science & Business Media, 2012.

[55] Dragan Gamberger and Nada Lavrac. "Expert-guided subgroup discovery: Methodology and application". In: *Journal of Artificial Intelligence Research* 17 (2002), pp. 501–527.

[56] Ning Gao et al. "Carotid intima-media thickness in patients with subclinical hypothyroidism: a meta-analysis". In: *Atherosclerosis* 227.1 (2013), pp. 18–25.

[57] E Geissner. "The Pain Perception Scale–a differentiated and change-sensitive scale for assessing chronic and acute pain". In: *Die Rehabilitation* 34.4 (1995), pp. 35–43.

[58] David Gilbert. *JFreeChart (Free Java class library for creating charts)*. 2005-2020. URL: http://www.jfree.org/jfreechart/.

[59] Sylvia Glaßer et al. "Reconstruction of 3D Surface Meshes for Bood Flow Simulations of Intracranial Aneurysms". In: *Proc. of the Annual Meeting of the German Society of Computer- and Robot-Assisted Surgery*. 2015, pp. 163–168.

[60] G Goebel and W Hiller. "Psychische Beschwerden bei chronischem Tinnitus: Erprobung und Evaluation des Tinnitus-Fragebogens (TF)". In: *Verhaltenstherapie* 2.1 (1992), pp. 13–22.

[61] Gerhard Goebel and Wolfgang Hiller. *Tinnitus-Fragebogen:(TF); ein Instrument zur Erfassung von Belastung und Schweregrad bei Tinnitus; Handanweisung*. hogrefe, Verlag für Psychologie, 1998.

[62] Alex Goldstein et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". In: *Journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65. DOI: 10.1080/10618600.2014.907095.

[63] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[64] J. C. Gower. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis". In: *Biometrika* 53.3/4 (Dec. 1966), p. 325. DOI: 10.2307/2333639.

[65] E. W. Gregg et al. "Changes in diabetes-related complications in the United States, 1990-2010". In: *N Engl J Med* 370.16 (2014), pp. 1514–23. ISSN: 1533-4406 (Electronic) 0028-4793 (Linking). DOI: 10.1056/NEJMoa1310799. URL: https://www.ncbi.nlm.nih.gov/pubmed/24738668.

[66] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42. DOI: 10.1145/3236009.

[67] R Gutekunst et al. "Ultrasonic diagnosis of the thyroid gland". In: *Deutsche medizinische Wochenschrift* 113.27 (1988), pp. 1109–1112.

[68] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of Machine Learning Research* 3.Mar (2003), pp. 1157–1182.

[69] Jonathan BS Halford and Stewart D Anderson. "Anxiety and depression in tinnitus sufferers". In: *Journal of psychosomatic research* 35.4-5 (1991), pp. 383–390.

[70] Mark Hall et al. "The WEKA data mining software: an update". In: *ACM SIGKDD expl. newsl.* 11.1 (2009), pp. 10–18.

[71] Mark A. Hall. "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning". In: *Proc. of International Conference on Machine Learning (ICML)*. 2000, pp. 359–366. ISBN: 1-55860-707-2. URL: http://dl.acm.org/citation.cfm?id=645529.657793.

[72] Mark A. Hall. "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning". In: *Proc. of International Conference on Machine Learning (ICML)*. 2000, pp. 359–366. ISBN: 1-55860-707-2. URL: http://dl.acm.org/citation.cfm?id=645529.657793.

[73] Jiawei Han, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". In: *ACM Sigmod Record*. Vol. 29. 2. ACM. 2000, pp. 1–12.

[74] John A Hartigan. "Printer graphics for clustering". In: *Journal of Statistical Computation and Simulation* 4.3 (1975), pp. 187–213. DOI: 10.1080/00949657508810123.

[75] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009. URL: https://web.stanford.edu/~hastie/ElemStatLearn/.

[76] Martin Hautzinger and Maja Bailer. "ADS-Allgemeine Depressionsskala". In: *Diagnostische Verfahren in der Psychotherapie*. Beltz, 2003.

[77] K. Hechenbichler and K. Schliep. "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification". In: *SFB 386, Ludwig-Maximilians University, Munich*. Vol. 399. sfb386. 2004. URL: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9.

[78] JL Henry and PH Wilson. "The Psychological Management of Tinnitus: Comparison of a Combined Cognitive Educational Program, Education Alone and a Waiting-List Control." In: *The International Tinnitus Journal* 2 (1996), pp. 9–20.

[79] Franciso Herrera et al. "An overview on subgroup discovery: foundations and applications". In: *Knowledge and information systems* 29.3 (2011), pp. 495–525.

[80] Tommy Hielscher et al. "A Framework for Expert-Driven Subpopulation Discovery and Evaluation Using Subspace Clustering for Epidemiological Data ". In: *Expert Systems with Applications* 113 (2018), 147–160. https://doi.org/10.1016/j.eswa.2018.07.003. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.07.003. URL: https://doi.org/10.1016/j.eswa.2018.07.003.

[81] Tommy Hielscher et al. "A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data". In: *Expert Systems with Applications* 113 (2018), pp. 147–160.

[82] Tommy Hielscher et al. "Identifying Relevant Features for a Multi-factorial Disorder with Constraint-Based Subspace Clustering". In: *Proceedings of the 29th IEEE Symposium on Computer-Based Medical Systems (CBMS)*. Dublin, Ireland: IEEE, 2016, pp. 207–212.

[83] Tommy Hielscher et al. "Mining longitudinal epidemiological data to understand a reversible disorder". In: *Proc. of Symposium on Intelligent Data Analysis*. 2014, pp. 120–130.

[84] Tommy Hielscher et al. "Using Participant Similarity for the Classification of Epidemiological Data on Hepatic Steatosis". In: *Proc. of the 27th IEEE Int. Symposium on Computer-Based Medical Systems (CBMS'14)*. accepted 03/2014, to appear. Mount Sinai, NY: IEEE, 2014.

[85] Aroon D Hingorani et al. "Prognosis research strategy (PROGRESS) 4: Stratified medicine research". In: *BMJ: British Medical Journal* 346.e5793 (2013), 9 pages. DOI: http://dx.doi.org/10.1136/bmj.e5793.

[86] Jerry L Hintze and Ray D Nelson. "Violin Plots: A Box Plot-Density Trace Synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184.

[87] Jonathan Hobson, Edward Chisholm, and Amr El Refaie. "Sound therapy (masking) in the management of tinnitus in adults". In: *Cochrane Database of Systematic Reviews* 11 (2012). DOI: 10.1002/14651858.CD006371.pub3.

[88] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[89] M Hörhold, BF Klapp, and U Schimmack. "Testungen der Invarianz und der Hierarchie eines mehrdimensionalen Stimmungsmodells auf der Basis von Zweipunkterhebungen an Patienten-und Studentenstichproben". In: *Z med Psychol* 2.1 (1993), pp. 27–35.

[90] M Hörhold et al. "Testing a screening strategy for identifying psychosomatic patients in gynecologic practice". In: *Psychotherapie, Psychosomatik, medizinische Psychologie* 47.5 (1997), pp. 156–162.

[91] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6 (1933), p. 417. DOI: 10.1037/h0071325.

[92] Jean-François Im, Michael J McGuffin, and Rock Leung. "GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2606–2614. DOI: 10.1109/TVCG.2013.160.

[93] Tsunenori Ishioka. "An expansion of X-means for automatically determining the optimal number of clusters". In: *Proceedings of International Conference on Computational Intelligence*. Vol. 2. 2005, pp. 91–95.

[94] Till Ittermann et al. "Inverse Association Between Serum Free Thyroxine Levels and Hepatic Steatosis: Results From the Study of Health in Pomerania". In: *Thyroid* 22.6 (2012), pp. 568–574. DOI: http://dx.doi.org/10.1089/thy.2011.0279.

[95] Avais Jabbar et al. "Thyroid hormones and cardiovascular disease". In: *Nature Reviews Cardiology* 14.1 (2017), pp. 39–55.

[96] Zhenghui Gordon Jiang, Simon C Robson, and Zemin Yao. "Lipoprotein metabolism in nonalcoholic fatty liver disease". In: *Journal of biomedical research* 27.1 (2013), p. 1.

[97] Mohammad Kachuee et al. *Cost-Sensitive Diagnosis and Learning Leveraging Public Health Data*. 2019. arXiv: 1902.07102 [cs.LG]. URL: https://arxiv.org/abs/1902.07102.

[98] Jalil Kazemitabar et al. "Variable Importance Using Decision Trees". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017, pp. 426–435. URL: https://proceedings.neurips.cc/paper/2017/file/5737c6ec2e0716f3d8a7a5c4e0de0d9a-Paper.pdf.

[99] Tanya Keenan et al. "Relation of uric acid to serum levels of high-sensitivity C-reactive protein, triglycerides, and high-density lipoprotein cholesterol and to hepatic steatosis". In: *The American journal of cardiology* 110.12 (2012), pp. 1787–1792.

[100] R C Kessler et al. "Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports". In: *Molecular Psychiatry* 21.10 (2016), pp. 1366–1371.

[101] Kenji Kira, Larry A Rendell, et al. "The feature selection problem: Traditional methods and a new algorithm". In: *AAAI*. Vol. 2. 1992, pp. 129–134.

[102] Paul Klemm et al. "3D Regression Heat Map Analysis of Population Study Data". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2015), pp. 81–90. DOI: 10.1109/TVCG.2015.2468291. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7194847.

[103] Paul Klemm et al. "Interactive visual analysis of image-centric cohort study data". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1673–1682.

[104] Willi Klösgen and Michael May. "Spatial subgroup mining integrated in an object-relational spatial database". In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer. 2002, pp. 275–286.

[105] Josua Krause, Adam Perer, and Kenney Ng. "Interacting with predictions: Visual inspection of black-box machine learning models". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 5686–5697.

[106] B Kröner-Herwig et al. "The management of chronic tinnitus: Comparison of an outpatient cognitive–behavioral group training to minimal-contact interventions". In: *Journal of psychosomatic research* 54.4 (2003), pp. 381–389. DOI: 10.1016/S0022-3999(02)00400-2.

[107] Max Kuhn. "Building Predictive Models in R Using the caret Package". In: *Journal of Statistical Software, Articles* 28.5 (2008), pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05. URL: https://www.jstatsoft.org/v028/i05.

[108] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[109] Max Kuhn and Ross Quinlan. *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.2. 2018. URL: https://CRAN.R-project.org/package=C50.

[110] Jens-Peter Kühn et al. "Noninvasive quantification of hepatic fat content using three-echo dixon magnetic resonance imaging with correction for T2∗ relaxation effects". In: *Investive Radiology* 46.12 (2011), pp. 783–789. DOI: http://dx.doi.org/10.1097/RLI.0b013e31822b124c.

[111] J.; Kumari M. J.; Subash and S. Jagdish. "How to Prevent Amputation in Diabetic Patients." In: *International Journal of Nursing Education.* 6 (2014), pp. 40–44.

[112] Michael Landgrebe et al. "The Tinnitus Research Initiative (TRI) database: a new approach for delineation of tinnitus subtypes and generation of predictors for treatment outcome". In: *BMC Medical Informatics and Decision Making* 10.1 (2010), p. 42. DOI: 10.1186/1472-6947-10-42.

[113] Berthold Langguth et al. "Different Patterns of Hearing Loss among Tinnitus Patients: A Latent Class Analysis of a Large Sample". In: *Frontiers in Neurology* 8 (2017), p. 46. DOI: 10.3389/fneur.2017.00046.

[114] Berthold Langguth et al. "Tinnitus and depression". In: *The world journal of biological psychiatry* 12.7 (2011), pp. 489–500.

[115] Berthold Langguth et al. "Tinnitus and depression". In: *The world journal of biological psychiatry* 12.7 (2011), pp. 489–500.

[116] Cicero Augusto de Lara Pahins, Nivan Ferreira, and Joao Comba. "Real-Time Exploration of Large Spatiotemporal Datasets based on Order Statistics". In: *IEEE Transactions on Visualization and Computer Graphics* (2019). DOI: 10.1007/s00778-020-00625-6.

[117] Katharina Lau et al. "The association between fatty liver disease and blood pressure in a population-based prospective cohort study". In: *Journal of Hypertension* 28.9 (2010), pp. 1829–1835. DOI: http://dx.doi.org/10.1097/HJH.0b013e32833c211b.

[118] Alexandra Lauric, Merih I Baharoglu, and Adel M Malek. "Ruptured status discrimination performance of aspect ratio, height/width, and bottleneck factor is highly dependent on aneurysm sizing methodology". In: *Neurosurgery* 71.1 (2012), pp. 38–46.

[119] Nada Lavrač et al. "Subgroup discovery with CN2-SD". In: *Journal of Machine Learning Research* 5 (2004), pp. 153–188.

[120] Matthijs van Leeuwen and Arno Knobbe. "Diverse subgroup set discovery". In: *Data Mining and Knowledge Discovery* 25.2 (2012), pp. 208–242.

[121] Joffrey L. Leevy et al. "A survey on addressing high-class imbalance in big data". In: *Journal of Big Data* 5.1 (2018), p. 42.

[122] Stan Lipovetsky and Michael Conklin. "Analysis of regression in game theory approach". In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pp. 319–330.

[123] Javier Lizardi-Cervera et al. "Association among C-reactive protein, Fatty liver disease, and cardiovascular risk". In: *Digestive diseases and sciences* 52.9 (2007), pp. 2375–2379.

[124] Hanna M van Loo et al. "Major depressive disorder subtypes to predict long-term course". In: *Depression and anxiety* 31.9 (2014), pp. 765–777.

[125] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[126] Scott M Lundberg et al. "Explainable AI for Trees: From Local Explanations to Global Understanding". In: *arXiv preprint arXiv:1905.04610* (2019).

[127] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

[128] Iris HL Maes et al. "Tinnitus: A Cost Study". In: *Ear and Hearing* 34.4 (2013), pp. 508–514. DOI: 10.1097/aud.0b013e31827d113a.

[129] Marcello R. P. Markus et al. "Hepatic Steatosis Is Associated With Aortic Valve Sclerosis in the General Population: The Study of Health in Pomerania (SHIP)". In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 33.7 (2013), pp. 1690–1695. DOI: http://dx.doi.org/10.1161/ATVBAHA.112.300556.

[130] Pablo Martinez-Devesa et al. "Cognitive behavioural therapy for tinnitus". In: *Cochrane Database of Systematic Reviews* 1 (2007). DOI: 10.1002/14651858.CD005233.pub3.

[131] Adrian Mayorga and Michael Gleicher. "Splatterplots: Overcoming Overdraw in Scatter Plots". In: *IEEE Transactions on Visualization and Computer Graphics* 19.9 (2013), pp. 1526–1538. DOI: 10.1109/TVCG.2013.65.

[132] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018). URL: https://arxiv.org/abs/1802.03426.

[133] Christoph Molnar. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/. 2020.

[134] Thanh-Phuong Nguyen, Corrado Priami, and Laura Caberlotto. "Novel drug target identification for the treatment of dementia using multi-relational association mining". In: *Scientific reports* 5 (2015).

[135] Uli Niemann et al. "Can we classify the participants of a longitudinal epidemiological study from their previous evolution?" In: *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2015, pp. 121–126. DOI: 10.1109/CBMS.2015.12.

[136] Uli Niemann et al. "Interactive Medical Miner: Interactively exploring subpopulations in epidemiological datasets". In: *ECML PKDD 2014, Part III, LNCS 8726*. Springer, 2014, pp. 460–463. DOI: 10.1007/978-3-662-44845-8_35.

[137] Uli Niemann et al. "Phenotyping chronic tinnitus patients using self-report questionnaire data: cluster analysis and visual comparison". In: *Scientific Reports* 10.1 (2020), p. 16411. DOI: 10.1038/s41598-020-73402-8. URL: https://doi.org/10.1038/s41598-020-73402-8.

[138] Cicero AL Pahins et al. "COVIZ: A System for Visual Information and Exploration of Patient Cohorts". In: *Proceedings of the VLDB Endowment* 12.12 (2019), pp. 1822–1825. DOI: 10.14778/3352063.3352075.

[139] Mykola Pechenizkiy et al. "Heart Failure Hospitalization Prediction in Remote Patient Management Systems". In: *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2010, pp. 44–49. DOI: http://dx.doi.org/10.1109/CBMS.2010.6042612.

[140] Dan Pelleg, Andrew W Moore, et al. "X-means: Extending k-means with efficient estimation of the number of clusters." In: *ICML*. Vol. 1. 2000, pp. 727–734.

[141] John S Phillips and Don McFerran. "Tinnitus retraining therapy (TRT) for tinnitus". In: *Cochrane database of systematic reviews* 3 (2010). DOI: 10.1002/14651858.CD007330.pub2.

[142] F. Pinheiro et al. "Extracting association rules from liver cancer data using the FP-growth algorithm". In: *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*. 2013. DOI: http://dx.doi.org/10.1109/ICCABS.2013.6629208.

[143] Thierry Poynard et al. "Apolipoprotein AI and alcoholic liver disease". In: *Hepatology* 6.6 (1986), pp. 1391–1395.

[144] Ren Qing-dao-er-ji, Rui Pang, and Yue Chang. "An Improved HotSpot Algorithm and Its Application to Sandstorm Data in Inner Mongolia". In: *Mathematical Problems in Engineering* (2020). DOI: 10.1155/2020/4020723. URL: https://www.hindawi.com/journals/mpe/2020/4020723/.

[145] John R. Quinlan. "Learning with Continuous Classes". In: *5th Australian Joint Conference on Artificial Intelligence*. Vol. 92. 1992, pp. 343–348.

[146] Ross Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[147] Ross Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[148] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.r-project.org/.

[149] Lenore Sawyer Radloff. "The CES-D scale: a self-report depression scale for research in the general population". In: *Applied psychological measurement* 1.3 (1977), pp. 385–401.

[150] Dheeraj Raju et al. "Exploring factors associated with pressure ulcers: A data mining approach". In: *Int. J. of Nursing Studies* (2014). ISSN: 0020-7489. DOI: http://dx.doi.org/10.1016/j.ijnurstu.2014.08.002. URL: http://www.sciencedirect.com/science/article/pii/S0020748914002053.

[151] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier". In: *Proc. of ACM SIGKDD Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[152] Stephanie A Riolo et al. "Prevalence of depression by race/ethnicity: findings from the National Health and Nutrition Examination Survey III". In: *American journal of public health* 95.6 (2005), pp. 998–1000.

[153] Bernd Röhrig et al. "Types of Study in Medical Research". In: *Dtsch Arztebl International* 106.15 (2009), pp. 262–268. DOI: 10.3238/arztebl.2009.0262. eprint: https://www.aerzteblatt.de/pdf.asp?id=64227. URL: https://www.aerzteblatt.de/int/article.asp?id=64227.

[154] Sylvia Saalfeld et al. "Semi-automatic neck curve reconstruction for intracranial aneurysm rupture risk assessment based on morphological parameters". In: *Proc. of Computer assisted radiology and surgery (CARS)*. 2018, to appear.

[155] James W Salazar et al. "Depression in Patients with Tinnitus: A Systematic Review". In: *Otolaryngology–Head and Neck Surgery* (2019), p. 0194599819835178.

[156] Wojciech Samek et al. "Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond". In: *arXiv preprint arXiv:2003.07631* (2020).

[157] Martin Schecklmann et al. "Cluster analysis for identifying sub-types of tinnitus: A positron emission tomography and voxel-based morphometry study". In: *Brain Research* 1485 (2012), pp. 3–9. DOI: 10.1016/j.brainres.2012.05.013.

[158] Winfried Schlee et al. "Visualization of global disease burden for the optimization of patient management and treatment". In: *Frontiers in Medicine* 4 (2017), p. 86. DOI: 10.3389/fmed.2017.00086.

[159] Miro Schleicher et al. "ICE: Interactive Classification Rule Exploration on Epidemiological Data". In: *Proc. of the 30th IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*. 2017, pp. 606–611. DOI: 10.1109/CBMS.2017.127. URL: https://doi.org/10.1109/CBMS.2017.127.

[160] Gudrun Scholler, Herbert Fliege, and Burghard F Klapp. "Fragebogen zu Selbstwirksamkeit, Optimismus und Pessimismus". In: *Psychother Psychosom Med Psychol* 49.8 (1999), pp. 275–283.

[161] Gideon Schwarz et al. "Estimating the Dimension of a Model". In: *Annals of Statistics* 6.2 (1978), pp. 461–464. DOI: 10.1214/aos/1176344136.

[162] David W Scott. "On optimal and data-based histograms". In: *Biometrika* 66.3 (1979), pp. 605–610. DOI: 10.1093/biomet/66.3.605.

[163] Lloyd S Shapley. "A value for n-person games". In: *Contributions to the Theory of Games* 2.28 (1953), pp. 307–317.

[164] Zaigham Faraz Siddiqui et al. "Predicting the post-treatment recovery of the patients suffering from TBI". In: *Int. Conf. on Brain Informatics and Health (BIH'14)*. Warsaw, Poland, Aug. 2014.

[165] Judith D Singer and John B Willett. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.

[166] Nalini Singh, David G Armstrong, and Benjamin A Lipsky. "Preventing foot ulcers in patients with diabetes". In: *Jama* 293.2 (2005), pp. 217–228.

[167] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, et al. "Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study". In: *JAMA* 282.18 (1999), pp. 1737–1744. DOI: 10.1001/jama.282.18.1737.

[168] Felix Stickel et al. "Genetic variation in the PNPLA3 gene is associated with alcoholic liver injury in caucasians". In: *Hepatology* 53.1 (2011), pp. 86–95. DOI: http://dx.doi.org/10.1002/hep.24017.

[169] Erik Štrumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and information systems* 41.3 (2014), pp. 647–665.

[170] Jimeng Sun et al. "A System for Mining Temporal Physiological Data Streams for Advanced Prognostic Decision Support". In: *IEEE Int. Conf. on Data Mining (ICDM)*. 2010, pp. 1061–1066. DOI: 10.1109/ICDM.2010.102.

[171] Noboru Takamura et al. "Thyroid function is associated with carotid intima-media thickness in euthyroid subjects". In: *Atherosclerosis* 204.2 (2009), e77–e81.

[172] Pang-Ning Tan et al. *Introduction to Data Mining*. 2nd ed. Pearson, 2019.

[173] Giovanni Targher, Christopher P. Day, and Enzo Bonora. "Risk of Cardiovascular Disease in Patients with Nonalcoholic Fatty Liver Disease". In: *New England Journal of Medicine* 363.14 (2010), pp. 1341–1350. DOI: http://dx.doi.org/10.1056/NEJMra0912063.

[174] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13. 2018. URL: https://CRAN.R-project.org/package=rpart.

[175] Matthew S Thiese. "Observational and interventional study design types; an overview". In: *Biochemia medica: Biochemia medica* 24.2 (2014), pp. 199–210. DOI: 10.11613/BM.2014.022. URL: https://www.biochemia-medica.com/en/journal/24/2/10.11613/BM.2014.022/fullArticle.

[176] Krysta J Trevis, Neil M McLachlan, and Sarah J Wilson. "A systematic review and meta-analysis of psychological functioning in chronic tinnitus". In: *Clinical psychology review* 60 (2018), pp. 62–86.

[177] Karin Tritt et al. "Entwicklung des Fragebogens ICD-10-Symptom-Rating (ISR)". In: *Zeitschrift für psychosomatische Medizin und Psychotherapie* 54.4 (2008), pp. 409–418. DOI: 10.13109/zptm.2008.54.4.409.

[178] Richard Tyler et al. "Identifying tinnitus subgroups with cluster analysis". In: *American journal of audiology* 17 (2 2008), pp. 176–184.

[179] Edip Unal et al. "Association of subclinical hypothyroidism with dyslipidemia and increased carotid intima-media thickness in children". In: *Journal of clinical research in pediatric endocrinology* 9.2 (2017), p. 144.

[180] Ioannis Valavanis et al. "Derivation of Cancer Related Biomarkers from DNA Methylation Data from an Epidemiological Cohort". In: *Engineering Applications of Neural Networks*. Springer, 2013, pp. 249–256.

[181] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[182] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. Springer, 2002.

[183] Jennell C Vick et al. "Data-driven subclassification of speech sound disorders in preschool children". In: *Journal of Speech, Language, and Hearing Research* 57.6 (2014), pp. 2033–2050.

[184] Giorgio Visani, Enrico Bagli, and Federico Chesani. "OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms". In: *arXiv preprint arXiv:2006.05714* (2020).

[185] M. Volmer-Thole and R. Lobmann. "Neuropathy and Diabetic Foot Syndrome". In: *Int J Mol Sci* 17.6 (2016). ISSN: 1422-0067 (Electronic) 1422-0067 (Linking). DOI: 10.3390/ijms17060917. URL: https://www.ncbi.nlm.nih.gov/pubmed/27294922.

[186] Henry Völzke et al. "A new, accurate predictive model for incident hypertension". In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 31.11 (2013), pp. 2142–2150. DOI: http://dx.doi.org/10.1097/HJH.0b013e328364a16d.

[187] Henry Völzke et al. "Cohort Profile: The Study of Health in Pomerania". In: *International Journal of Epidemiology* 40.2 (2011), pp. 294–307. DOI: http://dx.doi.org/10.1093/ije/dyp394.

[188] Henry Völzke et al. "Cohort profile: The Study of Health in Pomerania". In: *International Journal of Epidemiology* 40 (2011), pp. 294–307. DOI: 10.1093/ije/dyp394.

[189] Henry Völzke et al. "Hepatic steatosis is associated with an increased risk of carotid atherosclerosis". In: *World journal of gastroenterology: WJG* 11.12 (2005), p. 1848.

[190] Henry Völzke et al. "Prevalence trends in lifestyle-related risk factors: two cross-sectional analyses with a total of 8728 participants from the study of health in pomerania from 1997 to 2001 and 2008 to 2012". In: *Deutsches Ärzteblatt International* 112.11 (2015), p. 185.

[191] A. H. Weinberger et al. "Trends in depression prevalence in the USA from 2005 to 2015: widening disparities in vulnerable groups". In: *Psychological Medicine* 48.8 (2018), pp. 1308–1315. DOI: 10.1017/S0033291717002781.

[192] Marieke JH Wermer et al. "Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics". In: *Stroke* 38.4 (2007), pp. 1404–1410.

[193] Mary A Whooley et al. "Case-finding instruments for depression: Two questions are as good as many". In: *Journal of General Internal Medicine* 12.7 (1997), pp. 439–445.

[194] G Wiesner and EK Bittner. "Life expectancy, potential years of life lost (PYLL), and avoidable mortality in an East/West comparison". In: *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* 47.3 (2004), pp. 266–278.

[195] D. Randall Wilson and Tony R. Martinez. "Improved Heterogeneous Distance Functions". In: *Journal of Artificial Intelligenece Research* 6.1 (1997), pp. 1–34.

[196] Krist Wongsuphasawat and David Gotz. "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2659–2668.

[197] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[198] Jianping Xiang et al. "Hemodynamic–morphologic discriminants for intracranial aneurysm rupture". In: *Stroke* 42.1 (2011), pp. 144–152.

[199] Guo Yu, Daniela Witten, and Jacob Bien. *Controlling Costs: Feature Selection on a Budget*. 2020. arXiv: 1910.03627 [stat.ME]. URL: https://arxiv.org/abs/1910.03627.

[200] Zhiyuan Zhang, David Gotz, and Adam Perer. "Iterative cohort analysis and exploration". In: *Information Visualization* 14.4 (2015), pp. 289–307.

[201] Chuanlei Zhanga and Ralph L. Kodell. "Subpopulation-specific confidence designation for more informative biomedical classification". In: *Artificial Intelligence in Medicine* 58.3 (2013), pp. 155–163. DOI: http://dx.doi.org/10.1016/j.artmed.2013.04.008.