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

Uli Niemann

08.03.2021

Contents

Abstract

Subpopulation discovery is an essential objective of data analysis in medical research and contributes to the prevention and treatment of adverse medical conditions. Characteristic subpopulations are detected, for example, by identifying long-term determinants of diseases or by revealing patient subgroups with differential responses to treatment.

Traditional medical data analysis has been mostly hypothesis-driven. With the increasing volume and heterogeneity of medical data, these workflows are becoming increasingly impractical, as important relationships between variables may go undetected. Besides, medical studies often involve measurements that are collected repeatedly over time. Investigating hidden temporal information can potentially lead to new insights. While machine learning has the potential of automatically detecting previously unknown subpopulations, the results of complex black-box models must be made understandable. Therefore, the medical expert must be equipped with tools to understand, explore, and visualize the models, breaking down individual patterns to extract actionable insights.

This thesis proposes machine learning-based solutions for expert-driven subpopulation discovery in high-dimensional timestamped medical data.

The first part presents workflows to detect *comprehensible* and *distinct* subpopulations described by classification rules and clusters. We present novel visualizations and interactive tools to inspect and juxtapose the high-dimensional subpopulations and compare their change over time.

The second part covers workflows to exploit temporal information. We present a framework to extract *evolution features* that characterize the subpopulations' change over time. Furthermore, we provide a method to build representations from short temporal sequences.

The third part addresses the topic of *post-hoc interpretation* of complex black-box models. We propose an end-to-end data analysis workflow that includes steps for data augmentation, modeling, nesting model training with feature elimination, and post-hoc analysis of the trained

models. This workflow returns statistics and visualizations representing global feature importance, instance-individual feature importance, and subpopulation-specific feature importance for any machine learning model. Besides, we provide a solution for visualizing differences between *two* a priori defined subpopulations.

The proposed methods are evaluated with datasets from four medical studies:

- a longitudinal population study,
- an observational therapy study with data on self-report questionnaire responses from tinnitus patients,
- a clinical experiment with timestamped plantar pressure and temperature recordings from diabetes patients and healthy volunteers, and
- a retrospective clinical study with image data on intracranial aneurysms.

Zusammenfassung

Die *Entdeckung von Subpopulationen* stellt ein wesentliches Ziel der Datenanalyse in der medizinischen Forschung dar und trägt zur Vorbeugung und Behandlung von Erkrankungen bei. Charakteristische Subpopulationen werden beispielsweise durch die Identifizierung von Langzeitdeterminanten von Krankheiten oder durch die Bestimmung von Patientensubgruppen mit differenziellem Ansprechen auf eine Behandlung entdeckt.

Die traditionelle medizinische Datenanalyse war bisher überwiegend hypothesengetrieben. Mit der zunehmenden Menge und Heterogenität medizinischer Daten werden diese Workflows zunehmend ungeeignet, da wichtige Beziehungen zwischen Variablen unentdeckt bleiben können. Außerdem beinhalten medizinische Studien oft Messungen, die im Laufe der Zeit wiederholt erhoben werden. Die Untersuchung von verborgenen zeitlichen Informationen kann potenziell zu neuen Erkenntnissen führen. Während maschinelles Lernen das Potenzial hat, bisher unbekannte Subpopulationen automatisch zu erkennen, müssen die Ergebnisse komplexer Black-Box-Modelle verständlich gemacht werden. Dies erfordert, medizinische Expertinnen und Experten mit Werkzeugen auszustatten, die es ihnen ermöglichen, die Modelle zu interpretieren, zu explorieren und zu visualisieren, um individuelle Muster aufzuschlüsseln und daraus handlungsrelevante Erkenntnisse zu gewinnen.

In dieser Arbeit werden auf maschinellem Lernen basierende Lösungen für die expertengesteuerte Entdeckung von Subpopulationen in hochdimensionalen, zeitgestempelten medizinischen Daten vorgeschlagen.

Der erste Teil stellt Workflows vor, um *verständliche* und *unterscheidbare* Subpopulationen zu erkennen, die durch Klassifikationsregeln und Cluster beschrieben werden. Wir stellen neuartige Visualisierungen und interaktive Werkzeuge vor, um die hochdimensionalen Subpopulationen zu inspizieren und gegenüberzustellen sowie ihre Veränderung über die Zeit zu vergleichen.

Der zweite Teil befasst sich mit Workflows zur Modellierung zeitlicher Informationen. Wir stellen ein Framework zur Extrahierung von *Evolutionsvariablen* vor, die die zeitliche Veränderung der Subpopulationen beschreiben. Außerdem wird ein Verfahren zur Erstellung von Repräsentationen aus kurzen zeitlichen Sequenzen vorgestellt.

Der dritte Teil befasst sich mit dem Thema der *Post-hoc-Interpretation* von komplexen Black-Box-Modellen. Wir stellen einen Ende-zu-Ende-Datenanalyse-Workflow vor, der Schritte zur Datenanreicherung, Modellierung, Verzahnung von Modelltraining mit Variablen-Eliminierung und Post-hoc-Analyse der trainierten Modelle umfasst. Dieser Workflow liefert Kenngrößen und Visualisierungen, die die globale, instanz-individuelle und subpopulationsspezifische Variablenbedeutsamkeit für ein maschinelles Lernmodell jedweden Typs darstellen. Außerdem wird

eine Visualisierung von Unterschieden zwischen *zwei* apriorisch definierten Subpopulationen präsentiert.

Die vorgeschlagenen Methoden werden anhand von Datensätzen aus vier medizinischen Studien evaluiert:

- eine longitudinale Bevölkerungsstudie,
- eine beobachtende Therapiestudie mit Daten zu Selbstbeurteilungsfragebögen von Tinnitus-Patienten,
- ein klinisches Experiment mit zeitgestempelten Plantardruck- und Temperaturaufzeichnungen von Diabetes-Patienten und gesunden Probanden, und
- eine retrospektive klinische Studie mit Bilddaten zu intrakraniellen Aneurysmen.

Chapter 1

Introduction

1.1 Motivation and Objectives

Subpopulation discovery is an essential objective of data analysis in medical research [241, 247]. Knowledge of characteristic subpopulations can improve prevention (public health) and treatment (clinical medicine) of adverse medical conditions. Subpopulations are detected by (i) identifying long-term determinants and protective factors of a medical condition of interest [106, 30, 206], (ii) revealing subcohorts with increased disease prevalence or with different treatment response [233, 50, 53], and (iii) generating robust statistical models that can explain relationships between one or more independent variables and the target variable [268, 137, 104]. For example, epidemiologists attempt to discover associations between specific *features* (e.g., descriptors of lifestyle and demographics) and a *target variable* (e.g., obesity) in cohort studies by collecting and analyzing extensive participant data obtained from questionnaires, medical examinations, laboratory analyses, and imaging [282, 143, 136, 135]. In studies with a longitudinal design, these measurements are collected repeatedly over time and contain hidden temporal information, the investigation of which can potentially lead to new insights.

To find associations between variables, medical researchers usually first carefully derive hypotheses from clinical practice, experimental studies, or extensive literature reviews to test them formally for statistical significance [156]. However, with the ever-increasing volume and heterogeneity of medical data [246], traditional hypothesis-driven workflows are becoming increasingly impractical, as, for this reason, some critical inherent associations between variables may go undetected [277]. Machine learning can improve medical research by discovering understandable descriptions of patient or study participant subpopulations similar in terms of the target variable and can thus be used to derive new hypotheses [83, 253].

The proliferation of medical machine learning applications is motivated, among others, by the desire to make automated use of the plethora of information collected about study subjects, but sometimes also by the ubiquity of deep learning success stories in the media [40]. However, the ease of creating complex data-driven models is no guarantee that insights can be effortlessly derived [48]. Most state-of-the-art machine learning algorithms such as deep neural networks [100] and gradient boosting machines [85] 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 [2, 42].

It is critical that the application expert, who is not a practitioner but a scientist working in a clinical or epidemiological setting, is equipped with tools to understand, explore, and

visualize the models [217, 274] so that they can drill down to specific individual patterns and gain actionable insights that ultimately contribute to the 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 each application scenario [56].

This work proposes methods that serve as intelligent assistance to medical researchers to analyze high-dimensional, timestamped medical data. Hence, the core research question of the thesis is:

> **RQ:** How to derive accurate yet understandable patterns for subpopulation discovery in high-dimensional timestamped medical data?

Before, during, and after the generation of machine learning models, several challenges must be overcome for the medical expert to derive actionable knowledge. We translate these challenges into the following three goals:

GOAL1: *Comprehensibility and distinctiveness of subpopulations:* The extracted models, including clusters, rules, and other patterns, must be made understandable; preferably, the model generation process must also be comprehensible. Furthermore, we must minimize redundancy, which negatively affects the perceived quality of the model. Our task is to extract, process, and display the most relevant patterns for expert-driven model exploration.

GOAL2: *Exploitation of time:* Hidden temporal information must be exploited. Medical scholars search for long-term determinants of severe diseases. Finding patterns from subject "evolution" can contribute to this goal.

GOAL3: *Post-hoc interpretation of complex black-box models:* If the discovered patterns are not intrinsically interpretable, methods are required to extract the most relevant subpopulations and present them intuitively to the application expert.

1.2 Structure and Contributions of This Thesis

This thesis presents solutions to support medical researchers for expert-driven subpopulation discovery in high-dimensional, timestamped medical data. Design decisions and developments were partly inspired by suggestions from the respective domain experts and cooperation partners, including three tinnitus experts, an epidemiologist with statistical expertise, and a diabetes expert.

The thesis is organized into three parts and ten chapters tackling the research question and challenges mentioned above. Part I covers methods for subpopulation discovery in high-dimensional data. Part II focuses specifically on temporal aspects of medical datasets and provides approaches that extract informative representations from timestamped data. Part III addresses the post-hoc analysis of machine learning models and includes solutions to derive model-, observation-, and subpopulation-level insights from otherwise "opaque" black-box models.

- Chapter 2 (*Medical Background and Datasets*) presents the medical background relevant to this thesis, a brief comparison of medical study types, and an overview of the medical studies used to validate the proposed methods.
- Chapter 3 (*Interactive Discovery and Inspection of Subpopulations*) presents a workflow for interactive data-driven analysis of population-based cohort data using hepatic steatosis as an example. It includes steps (i) to detect subpopulations that have different distributions with respect to the target variable, (ii) to classify each subpopulation taking class imbalance into account, and (iii) to detect variables associated with the outcome.

- Chapter 4 (*Identifying Distinct Subpopulations*) refines the analysis of the previous chapter by examining redundancy in large rule sets describing subpopulations. We present a workflow that extracts a smaller number of "representative" rules, i.e., rules that avoid instance overlap as much as possible, thus covering different subpopulations.
- Chapter 5 (*Visual Identification of Informative Features*) introduces a parameter-free clustering approach for deriving phenotypes, phenotype exploration, and visual juxtaposition of phenotypes in a high-dimensional feature space.
- Chapter 6 (*Constructing Evolution Features to Capture Change over Time*) presents a solution for cohort analysis in longitudinal cohort study data to construct "evolution features" from latent temporal information describing the cohort participants' change over time.
- Chapter 7 (*Feature Extraction from Short Temporal Sequences for Clustering*) complements the previous solution by presenting an approach to create representations from short temporal sequences via clustering in experimental data.
- Chapter 8 (*Post-Hoc Interpretation of Classification Models*) builds upon the insights of the previous chapters on the role of features in subpopulation understanding. We propose a method that makes already learned, complex classification models understandable to the domain experts. We combine the classification of high-dimensional medical data with model explanation using post-hoc interpretation methods. To this end, we use Shapely value explanations (SHAP), LASSO coefficients, and partial dependency plots. 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.
- Chapter 9 (*Subpopulation-Specific Learning and Post-Hoc Model Interpretation*) addresses the issue of visualizing differences between *two* subpopulations in temporal data. For this purpose, we derive a post-hoc interpretation measure to assess the difference in the predictors' association with the target variable between two subpopulations.
- Chapter 10 (*Summary and Future Work*) concludes the thesis by summarizing the contributions and providing a detailed outlook for the presented work.

For the validation of the proposed methods, we used datasets from the following epidemiological and clinical studies:

SHIP the longitudinal population "Study of Health in Pomerania" [282],

CHA an observational therapy study involving data on self-report questionnaire responses from tinnitus patients [204],

DIAB a clinical experiment yielding timestamped plantar pressure and temperature recordings from diabetes patients and non-diabetic volunteers [202], and

ANEUR a retrospective clinical study involving image data on intracranial aneurysms [203].

These datasets are used for method validation as follows:

Dataset	Chapter 3	Ch. 4	Ch. 5	Ch. 6	Ch. 7	Ch. 8	Ch. 9
SHIP	x	x		x			x
CHA			x			x	x
DIAB					x		
ANEUR						x	

Chapter 2

Medical Background and Datasets

Medical research on data from clinical and epidemiological studies lays the foundation for decisions about diagnosing and treating multifactorial conditions such as diseases and disorders. Major goals are to identify long-term determinants and protective factors for an outcome of interest [106, 30, 206], discover subpopulations with increased disease prevalence [233, 50, 53], and study intervention effects by generating statistical models explaining cause-and-effect relationships [268, 137, 104]. Traditional medical data analysis pipelines are usually structured in a *hypothesis-driven* way as follows [156]:

(1) A medical scientist formulates a hypothesis based on observations in clinical practice or current research. Possible examples include: "How does a risk factor such as alcohol abuse affect the prevalence of a particular outcome?", or "What effect does a novel therapy have on patients with depressive symptoms?"
(2) A small set of relevant variables that can be controlled for confounders is selected to test this hypothesis. Variable selection may include controlling for confounders. The data necessary to test the hypothesis are then collected.
(3) The strength of associations between the selected variables and the outcome is assessed using regression models and statistical methods.
(4) Based on the results, inferential statistical calculations will be performed, and conclusions will be drawn that may support the implementation of new preventive interventions or the use of appropriate treatments in high-risk patients.

However, with the advent of *big data* [207] in various fields, including medicine, data volume and heterogeneity are increasing dramatically, making traditional hypothesis-driven workflows increasingly inadequate as important relationships between variables may go undetected [277]. In this thesis, we present methods that deal with different aspects of high-dimensional timestamped medical data. We validate our methods on a variety of datasets from diverse study types.

This chapter is divided into two parts. Section 2.1 provides a brief comparison of medical study types. In Section 2.2, we present the studies and the data samples for which we developed the methods proposed in this thesis.

2.1 Brief Comparison of Medical Study Types

Primary medical research can be divided into basic, clinical, and epidemiological studies [229]. The following comparison of these study types is based on the reviews of Thiese [264] and Röhrig et al. [229] if not indicated otherwise.

Basic research. Basic medical research (or experimental research) aims to improve the understanding of cellular, molecular, and physiological mechanisms of human health and disease by conducting cellular and molecular investigations, animal studies, drug and material property studies in tightly controlled laboratory environments. To study the effects of one or more variables of interest on the outcome, all other variables are usually held constant, and only the variables of interest are varied. The carefully standardized experimental conditions of basic medical studies ensure high internal validity, but these conditions often cannot be easily transferred to clinical practice without compromising the results' generalizability.

Clinical studies. Clinical studies are generally classified into *interventional* (*experimental*) studies and *non-interventional* (*observational*) studies. The general objective of an intervention study is to compare different treatments within a patient population whose members differ as little as possible except for the treatment arm. A common example is a pharmaceutical study that aims to validate the efficacy and safety by investigating or establishing a drug's main and side effects, absorption, metabolism, and excretion. Selection bias can be avoided by appropriate measures, in particular by randomly assigning patients to groups. Treatment may be a medication, surgery, therapeutic use of a medical device (e.g., a stent), physical therapy, acupuncture, psychosocial intervention, rehabilitation, training form, or diet. A *randomized controlled trial* (RCT) is considered the gold standard of study design [116]. Selection bias is minimized by (a) randomly assigning patients to treatment and control groups and (b) ensuring equal distribution of known and unknown influencing variables (confounders), such as risk factors, comorbidities, and genetic variability. RCTs are thus suitable for obtaining an unambiguous answer to a clear question concerning the (causal) efficacy of a treatment.

Non-interventional clinical trials are patient-based observational studies in which patients either receive an individually defined treatment, or all patients receive the same treatment. An example of a non-interventional design is a study investigating the regular use of drugs in therapies. Here, treatment, diagnosis, and monitoring do not follow a predefined study protocol but rather medical practice alone. Data analysis is often retrospective. Whether a study design is *prospective* or *retrospective* depends on the sequence of hypothesis generation and data collection. In prospective studies, hypothesis generation comes before data collection. First, the hypotheses to be tested are defined, e.g., regarding a new treatment procedure. Then, data are collected specifically for hypothesis testing. By first formulating testable hypotheses, it is possible to ensure that the research questions can actually be answered with the measured data. A retrospective study design means that data collection took place before the study began.

Epidemiological studies. Epidemiological studies are usually interested in the distribution and change over time of the incidence of diseases and their causes in the general population and subpopulations. *Cohort studies* examine individuals, some of whom do not have the health outcomes of interest at the beginning of the observation period and assess exposure status to various health-related conditions [96]. The included subjects are then followed up over time in *longitudinal* studies (as opposed to *cross-sectional* studies), where the outcomes of interest are recorded in multiple *waves*. With these data, researchers can establish subgroups of subjects by exposure status, sort them by exposure, and compare the incidence or prevalence of a disease

among exposure categories. 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.

2.2 Datasets Investigated in This Thesis

In this section, we present the datasets and the associated studies investigated in this thesis, namely

- the Study of Health in Pomerania in Section 2.2.1,
- an observational therapy study on the health of tinnitus patients at baseline and after therapy in Section 2.2.2,
- a clinical experiment study on diabetic foot syndrome in Section 2.2.3, and
- a retrospective clinical study with image data on intracranial aneurysms in Section 2.2.4.

2.2.1 The Study of Health in Pomerania (SHIP)

After the reunification of Germany, it was found that life expectancy was significantly lower in the East than in the West [286]. Furthermore, there were regional differences within the new states, with the lowest life expectancy found in the Northeast [286, 293]. To find causal risk factors of mortality and other conditions in the northeastern German population, the Community Medicine Research Center in Greifswald established the *Study of Health in Pomerania* (SHIP) [282], a longitudinal epidemiological study of two independent cohorts in northeastern Germany. SHIP seeks to describe a broad spectrum of health conditions rather than focusing on a specific target disease [282]. In particular, major study objectives include investigations of the prevalence of common diseases and their risk factors, the correlation and interaction between risk factors and diseases, the progression from subclinical to manifest diseases, the identification of subpopulations with increased health risk, the prediction of concomitant diseases, as well as the usage and costs of medical service.

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 to 6 years) examination program that encompasses personal interviews, body measurements, exercise electrocardiogram, laboratory analysis, ultrasound examinations, and whole-body magnetic resonance tomography (MRT). Baseline examinations for the first cohort were performed between 1997 and 2001 (SHIP-0, N = 4308). Followup examinations were carried out in 2002-2006 (SHIP-1, n = 3300), 2008-2012 (SHIP-2, n = 2333), 2014 - 2016 (SHIP-3, n = 1718) and since 2019 (SHIP-4). Figure 2.1 illustrates participant response and age distribution of show-ups across study waves. 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.

The examination program changed across waves. For example, MRT was not performed until SHIP-2; liver ultrasound was performed in SHIP-0 and SHIP-2 but not in SHIP-1; dermatologic examinations were performed in SHIP-1 and SHIP-2 but not in SHIP-0.

Our analyses focus on the disorder *hepatic steatosis*, also known as “fatty liver”, characterized by a high accumulation of fat in the liver, occurring in approximately 30% of all adults [282, 284]. Risk factors include alcohol abuse, obesity, metabolic syndrome, and diabetes [281]. Liver

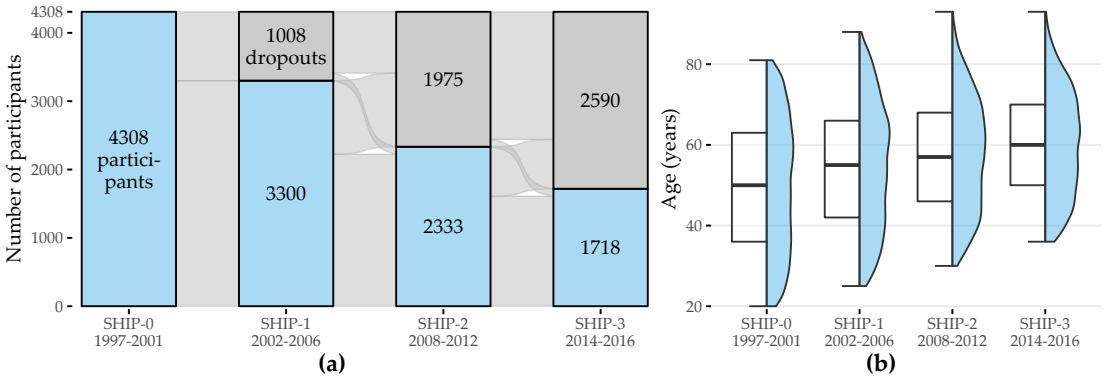


Figure 2.1: **Participation response and age distribution of show-ups across SHIP study waves.** (a) Change in the number of show-ups and non-respondents relative to the cohort size across waves. (b) Age distribution of show-ups for each wave.

biopsy is considered the diagnostic gold standard [7] but is associated with intermediate risk for the patient. Non-invasive diagnostic techniques include MRI, CT, and ultrasound. Because hepatic steatosis is usually asymptomatic, it often goes undetected, which can develop into more serious diseases, such as steatohepatitis, cirrhosis, hepatocellular carcinoma, or even liver failure [7].

We used SHIP data subsets to validate the methods presented in Chapters 3, 4, 6, and 9.

2.2.2 The Charité Tinnitus Patients Observational Therapy Study Dataset (CHA)

Tinnitus is the perception of a phantom sound in the absence of an external sound source. It is a complex, multifactorially caused and maintained phenomenon and is estimated to affect 10% to 15% of the adult population [17]. The associated annual economic burden is 19.4 billion USD in the United States [28] and 6.8 billion EUR in the Netherlands alone [186]. Clinical evaluation of tinnitus is challenging due to patient heterogeneity in tinnitus perception (laterality, pitch, noise characteristics, frequency, duration, chronicity), risk factors (including hearing loss, temporomandibular joint disorder, aging), comorbidities (including hyperacusis, depression, sleep disorders), perceived distress, and treatment response [45]. These differences complicate the identification of an appropriate and effective treatment modality. Currently, there is no treatment gold standard: sound therapy (masking), informational counseling (minimal contact education), cognitive behavioral therapy, and tinnitus retraining have been shown to be effective for some patients, but there is also evidence that not all patients benefit equally from these forms of treatment [133, 160, 121, 188, 212]. Due to the heterogeneous nature of the tinnitus symptom and the unclear evidence base regarding its treatment and management, identification of patient subgroups is critical to stratify individual pathophysiology and treatment pathways [167, 269, 166].

The *"Charité tinnitus patients observational therapy study dataset" (CHA)* includes self-report data from 4103 tinnitus patients treated at the Tinnitus Center of Charité Universitätsmedizin Berlin, Germany, between January 2011 and October 2015. All patients were 18 years of age or older and had suffered from tinnitus for at least 3 months. Exclusion criteria were the presence of acute psychotic illness or addiction, deafness, and insufficient knowledge of the German language. Treatment included a 7-day multimodal program with intensive and daily

informational counseling, detailed ear-nose-throat psychological diagnosis, cognitive behavioral therapy interventions, hearing exercises, progressive muscle relaxation, and physical therapy. At baseline (*T0*; before the start of therapy) and after treatment (*T1*), patients were asked to complete several self-report questionnaires. These questionnaires were selected to obtain a comprehensive assessment of tinnitus, including tinnitus-related distress and the psychosomatic background of tinnitus with anxiety, depression, the general quality of life, and experienced physical impairments.

Table 2.1 provides an overview of all questionnaires used in our analyses. Most questionnaires contain multiple-choice items with answers on a Likert scale. For example, the "Tinnitus Questionnaire" [98] (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" (coded as 0), "partly true" (1), and "true" (2). Some questionnaires also include aggregate variables called "subscales" and "total scores." For example, the TQ total score (TQ_distress) is calculated as the sum of 40 item values, with 2 items used twice [98], resulting in a range of values from 0 to 84, with higher values representing higher tinnitus-related distress. The cutoff value 46 [98] is used to distinguish between *compensated* (0-46) and *decompensated* (47-84) tinnitus. Furthermore, the average time to answer an item was recorded for each questionnaire. Figure 2.2 provides a graphical representation of demographic data for 3803 (92.7%) patients with complete data for the socio-demographics questionnaire [36] (SOZK) and TQ score.

Table 2.1: **Description of questionnaires that form the basis for CHA.** |F|: total number of items, subscales and total scales.

	Name	**Scope**	**|F|**
1	ACSA: Anamnestic Comparative Self-Assessment [26]	Quality of life	1
2	ADSL: General Depression Scale [222, 119]	Depressive symptoms	22
3	BI: Berlin Complaint Inventory [139]	General well-being, autonomic nervous system, pain and emotionality	29
4	BSF: Berlin Mood Questionnaire [138]	Mood	36
5	ISR: ICD-10 Symptom Rating [267]	Mental disorders	36
6	PHQK: (Short-form) Patient Health Questionnaire [252]	Symptoms of depression and anxiety	16
7	PSQ: Perceived Stress Questionnaire [80]	Stress	35
8	SES: Pain Perception Scale [90]	Pain	29
9	SSKAL: Visual Analog Scales Pain	Pain impairment, frequency and intensity	3
10	SF8: Short Form 8 Health Survey [37]	Health-related quality of life	18
11	SOZK: A socio-demographics questionnaire [36]	Gender, partnership status, education, employment status, among others	27
12	SWOP: Self-Efficacy-Optimism-Pessimism Scale questionnaire [240]	Self efficacy, optimism, pessimism	12
13	TINSKAL: Visual analogue scales	Tinnitus loudness, frequency and distress	3
14	TLQ: Tinnitus Localization and Quality questionnaire [97]	Location (left, right, bilateral, entire head) and sound of tinnitus	8

	Name	Scope	|F|
15	TQ: Tinnitus Questionnaire (German version) [98]	Tinnitus-related distress and tinnitus severity	60

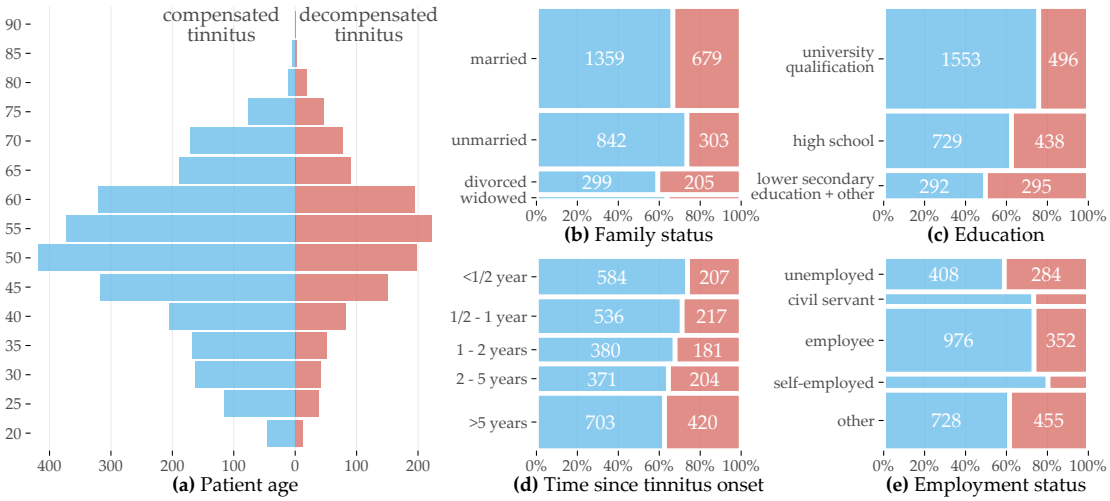


Figure 2.2: **Patient demographics (CHA).** Overview of patient demographics by degree of tinnitus distress measured before therapy commencement.

We used the CHA dataset to validate the methods developed in Chapters 5, 8, and 9.

2.2.3 The Diabetic Foot Clinical Experiment Study (DIAB)

Diabetic foot syndrome is an umbrella term for foot-related problems in diabetic patients. Up to one in four diabetes patients will develop a foot ulcer during their lifetime [32], with many at risk of amputations in the next four years [165]. More than 85% of foot amputations are due to foot ulcers [102, 280]. The rate of foot amputations in diabetes patients is estimated to be 17-40 times higher than in the general population [73]. DFS patients are predisposed to peripheral sensory neuropathy, which results, for example, in patients being unaware of the temperature of their feet or the pressure applied to them. Affected individuals may even injure themselves without realizing it. Excessive plantar pressures can exacerbate tissue destruction and increase the lifetime risk of foot ulceration [250]. However, understanding of the pathomechanisms underlying tissue destruction in the absence of trauma is limited.

At the university hospital of Magdeburg, Germany, 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 surrogate of tissue perfusion. For this purpose, plantar pressure and temperature changes in the feet were recorded during extended episodes of standing. Custom-made shoe insoles [103] equipped with eight temperature sensors and eight pressure sensors at preselected positions were used for data acquisition (Figure 2.3 (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 person's body temperature and is also affected by the environmental temper-

ature. 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".

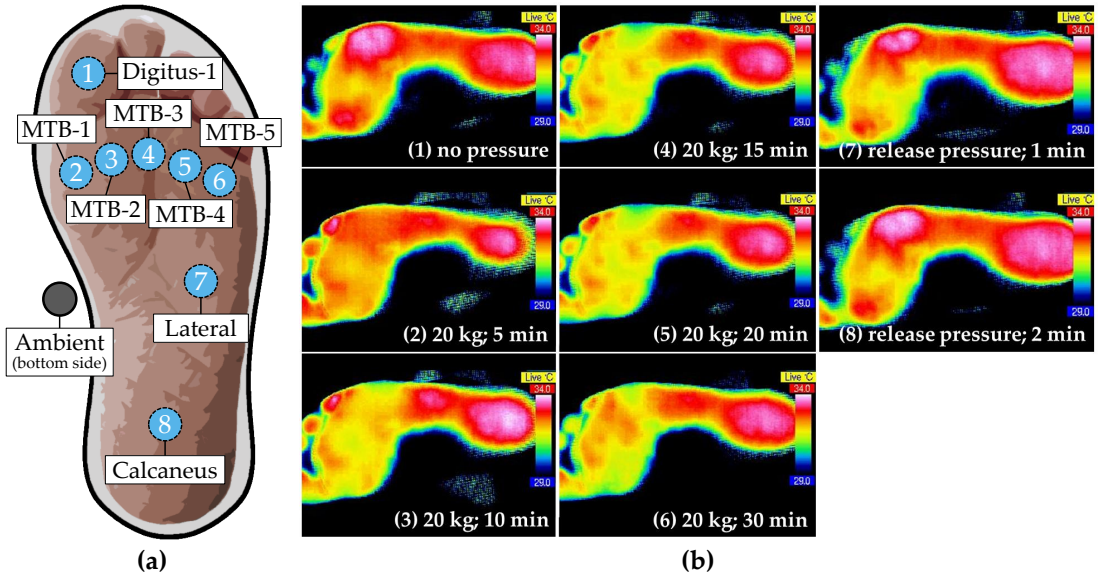


Figure 2.3: **Positions of pressure- and temperature sensors on insole and temporal, pressure-dependent temperature change.** (a) Sensor positioning on the insole in relation to foot placement. (b) Thermographic infrared images showing a healthy subject in a seated position with no pressure applied to the feet (before) and after placement of a 20 kg weight on both thighs. The measured temperature ranged from 29°C (blue) to 34°C (red). A time-dependent temperature decrease was observed predominantly in the forefoot region, visualized by yellow color during pressure application. A rapid temperature increase was noted within 1 min after pressure relief. MTB: metatarsal bone. The figure is adapted from [202].

Data collection began immediately after the shoes were put on. Participants were asked to follow a predefined sequence of actions, i.e., alternating between standing (*stance episode*) and sitting (*pause*). A session consisted of 6 stance episodes lasting 5, 10, 20, 5, 10, and 20 minutes, respectively, separated by pause episodes lasting 5 minutes each. Participants were instructed to apply equal pressure to both feet while standing. Participants did not receive immediate feedback on the actual application of pressure during the sessions, but they were verbally encouraged by the study nurses to maintain pressure while standing without releasing it. In the seated position, participants were instructed to release pressure for 5 minutes while maintaining contact with the insole. Participants were explicitly asked to adhere to these instructions, i.e., not to release pressure during a standing episode temporarily. The study protocol further included that the measurements were performed twice, once at room temperature of approximately 22°C and once outdoors at an ambient temperature of approximately 16°C. The two measurements were performed on two independent days.

The thermographic images in Figure 2.3 (b) visualize exemplary changes in plantar temperature in a healthy subject in a sitting position before pressure application (1), after placing a 20 kg weight on the front of the thigh (2-6), and after removing the additional weight (7-8). During pressure application, a gradual temporal decrease in temperature was noted predominantly in the forefoot. After pressure relief, a rapid temperature increase was observed within 1 min.

We used the DIAB dataset to validate the methods developed in Chapter 7.

2.2.4 The Intracranial Aneurysm Angiography Image Dataset (ANEUR)

Intracranial aneurysms are pathologic dilations of the intracranial vessel wall, often in the form of a dilation. They bear a risk of rupture, leading 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, location, and patient age and sex [291]. Further studies identified parameters, such as aspect ratio, undulation index, and nonsphericity index, to be statistically significant with respect to aneurysm rupture status [62, 297]. However, although these studies allow for retrospective analysis, the clinician needs further guidance if an asymptomatic aneurysm (as an accidental finding) was detected and the rupture risk should be determined.

We developed methods for the retrospective *"Intracranial Aneurysm Angiography Image Dataset"* (ANEUR) 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.

Motivated by the results of Baharoglu et al. [18], who found differences between sidewall and bifurcation aneurysms (cf. Figure 2.4) in terms of the relationship of several morphological parameters and rupture status, we learn different models for the subset of sidewall aneurysms (9 (37.5%) of 24 ruptured) and the subset of bifurcation aneurysms (29 (46.8%) of 62 ruptured). Additionally, we run experiments on a combined group (43 of 100 ruptured) containing 14 additional samples that could not be clearly determined to be either sidewall or bifurcation aneurysms.

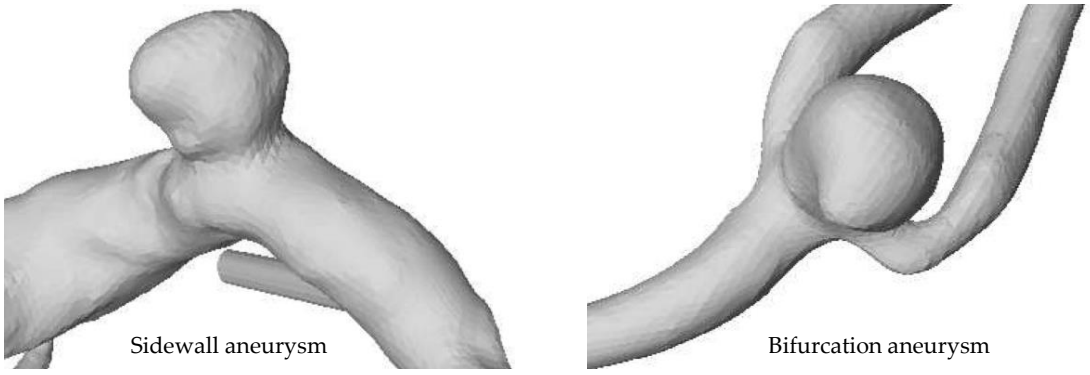


Figure 2.4: **Sidewall and bifurcation aneurysms.** Illustration of an aneurysm at the side of the parent vessel wall (left) and an aneurysm at a vessel bifurcation (right). The figure is adapted from [203].

We used the ANEUR dataset to validate the methods developed in Chapter 8.

Part I

Subpopulation Discovery in High-Dimensional Data

Chapter 3

Interactive Discovery and Inspection of Subpopulations

Brief Chapter Summary

Analysis of population-based cohort data has been mostly hypothesis-driven. We present a workflow and an interactive application for data-driven analysis of population-based cohort data using hepatic steatosis as an example. Our mining workflow includes steps

i. to discover subpopulations that have different distributions with respect to the target variable,
ii. to classify each subpopulation taking class imbalance into account, and
iii. to identify variables associated with the target variable.

We show that our workflow is suited (a) to build subpopulations before classification to reduce class imbalance and (b) to drill-down on the derived models to identify predictive variables and subpopulations worthy of further investigation.

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.
- Uli Niemann, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. "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.

This chapter is organized as follows. In Section 3.1, we motivate for classification and interactive subpopulation discovery in epidemiological cohort studies and review related work. We present our workflow and the interactive assistant in Section 3.2. In Section 3.3, we report our results and

main findings. The chapter closes with a summary and a discussion of the main contributions in Section 3.4.

3.1 Motivation and Comparison to Related Work

Medical decisions about the diagnosis and treatment of multifactorial conditions such as diseases and disorders are based on clinical and epidemiological studies [69]. The latter contain information about participants with and without a disease and allow learning of discriminatory models and, in longitudinal designs, understanding disease progression. For example, several studies identified risk factors (such as obesity and alcohol consumption) and comorbidities (such as cardiovascular disease) associated with hepatic steatosis [145, 171, 254, 260, 187]. However, these studies identified risk factors and associated outcomes that relate to the *entire* population. Our work arose from the need to identify such factors for subpopulations to promote personalized diagnosis and treatment, as expected in personalized medicine [130, 283].

3.1.1 Role of Subpopulations in Classifier Learning for Cohorts

Classification on subpopulations was studied by Zhanga and Kodell [302], who pointed out that classifier performance on the whole dataset may be low if the entire population is very heterogeneous. Therefore, they first trained an ensemble of classifiers and then used each ensemble member's predictions to create a new feature space. They performed hierarchical clustering to partition the instances into three subpopulations: one where the prediction accuracy is high, one where it is in the intermediate range, and one where it is low. Using this approach, Zhanga and Kodell partition the original data set into subpopulations that are easy or hard to classify. While the method seems appealing in general, it appears inappropriate for the three-class problem of the SHIP data, which has a highly skewed distribution, so it is clear that the low classification accuracy is caused (in part) by the class imbalance. Therefore, we exploratively examined the data set *before* classification to identify less skewed subpopulations and *after* classification to determine – within each subpopulation – variables strongly associated with the target.

Pinheiro et al. performed association rule discovery in patients with liver cancer [214]. The authors pointed out that early detection of liver cancer reduces the mortality rate. Early detection is still difficult because patients often do not show symptoms in the early stages of liver cancer [214]. Pinheiro et al. used the association rule algorithm FP-growth [113] to discover high-confidence association rules and high-confidence classification rules related to liver cancer mortality. We also considered association rules promising for medical data analysis because they are easy to compute and produce results that are understandable to humans. Therefore, we used association rules as the primary method, but for epidemiological data and classification rather than for mortality prediction. To use association rules for classification, we specified that the rule's consequence is the target variable.

3.1.2 Workflows for Expert-Machine Interaction for Cohort Construction and Analysis

Zhang et al. [301] addressed the increasing technical challenges of medical expert-driven subpopulation discovery due to increasingly large and complex medical data, often including information from hundreds of variables for thousands of patients in the form of tables, images, or text. In the past, it was sufficient for a physician to have some basic knowledge of statistics and spreadsheet software such as Microsoft Excel to analyze a small table of patient data. Today,

more effective and efficient approaches for managing, analyzing, and summarizing extensive medical data are available [301]. However, domain experts typically rely on technical experts to help them perform these tasks. This back-and-forth is often slow, tedious, and expensive. Therefore, it would be better to provide the domain expert with a technical tool that allows them to perform exploratory analysis by themselves quickly. Zhang et al. [301] presented CAVA, a system that includes various subgroup visualizations (called "views") and analytical components (called "analytics") for subgroup comparison. The main panel in Figure 3.1 shows one of the views: a flowchart [295] of patient subgroups with the same sequence of symptoms. The user can obtain additional summaries by interacting with the visualization, for example, by dragging and dropping one of the boxes in the flowchart onto one of the entries in the analysis panel. The user can also expand the selected cohort by having the tool search for patients who do not strictly meet the current inclusion criteria but are somewhat *similar* to the selected patient subpopulation of interest [68].

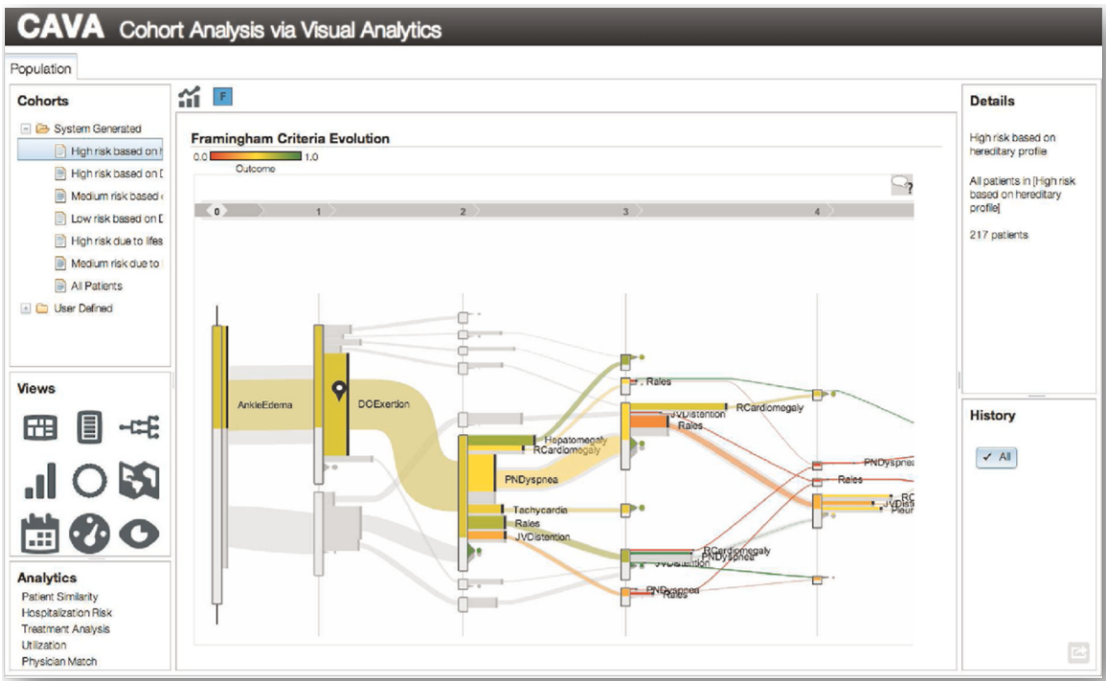


Figure 3.1: **CAVA's graphical user interface.** The flowchart visualizes subgroups of cardiac patients organized by the common occurrence of symptoms. Arc color represents the hospitalization risk. The user can switch between graphical representations and data processing methods by dragging and dropping. The upper right panel contains detailed information about the currently selected patients. The lower right panel contains a provenance graph that allows the user to undo operations and revisit previous interaction steps. The figure is taken from [301].

Krause et al. [159] argued that model selection should not be based only on global performance metrics such as accuracy, as these statistics do not contribute to a better understanding of the model's reasoning. Moreover, a complex but highly accurate model does not automatically guarantee actionable insights. Krause et al. propose Prospector [159], a system that provides diagnostic components for complex classification models based on the concepts of partial dependence (PD) plots [85]. PD plots are a popular tool for visualizing *marginal effects* of features on the predicted probability of the target. Briefly, each point on a PD plot represents the model's average prediction over all observations, assuming that these observations had a fixed value for a feature of interest. A feature whose PD curve exhibits a high range or high variability is

considered more influential on model prediction than a feature with a flat PD curve. Closely related to PD plots are individual conditional expectation (ICE) plots, [99] which display a curve for each observation, helping to reveal contrasting subpopulations that might "average out" in a PD plot. Prospector combines PD and ICE curves to show the relationship between a feature and model prediction at a (*global*) model level and a (*local*) patient-individual level. Besides, a custom color bar is provided as a more compact alternative to ICE curves (Figure 3.2 (a)). A stacked bar graph shows the distribution of predicted risk scores for each study group (Figure 3.2 (b)). The user can click on a specific decile to obtain a list of individual patients with their exact predicted score and label. In this way, patients whose prediction scores are close to the decision threshold can be further investigated. For each feature, the authors calculate the "most impactful feature change": given a patient's current feature values, they identify a near-counterfactual value that leads to a large change in the predicted risk score by minimizing the difference from the original feature value and maximizing the predicted risk score. The top 5 of these so-called "suggested changes" are displayed – separately for increasing and decreasing disease risk - in a table (cf. Figure 3.2 (c)) and integrated as interactive elements into the IC color bars (cf. Figure 3.2 (d)).

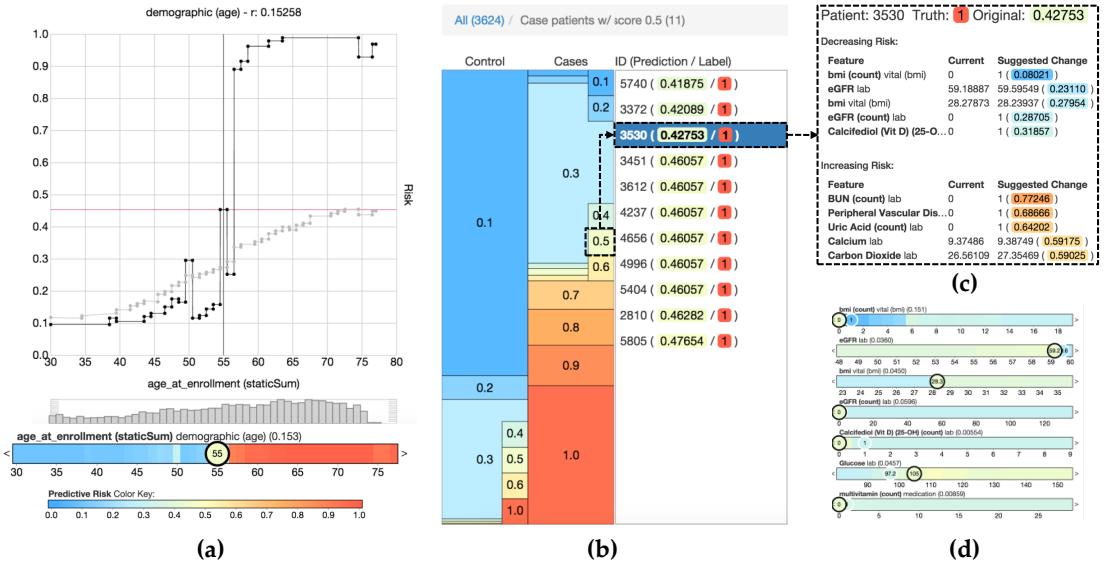


Figure 3.2: **Selected model diagnostics of Prospector.** (a) The upper plot shows two curves for the characteristic "age": the gray partial dependence (PD) curve represents the marginal prediction of the model over all patients, while the black individual conditional expectation (ICE) curve illustrates the effect of counterfactual ages on the predicted risk of diabetes for an example patient. The histogram shows the age distribution. The color bar below is a compact representation of the ICE curve above; the circled value represents the selected patient's feature value. (b) Stacked bars show the distribution of predicted risk scores for each study group. Clicking on one of the bars opens a table showing the ID, predicted risk, and true label for all patients belonging to the selected decile of predicted risk. (c) Summary table of "most impactful feature changes" for a decreasing (upper group) and an increasing (lower group) predicted risk: each row shows the actual feature value and the "suggested change," i.e., a similar but counterfactual value that would lead to a significant change in predicted risk. (d) Multiple PD color bars augmented with suggested changes (labels outlined in white). The figure is adapted from [159].

Pahins et al. [208] presented COVIZ, a system for cohort construction in large spatiotemporal

datasets. COVIZ includes components for exploratory data analysis of treatment pathways and event trajectories, visual cohort comparison, and visual querying. One of the design goals of COVIZ was to be fast, e.g., by using efficient data structures such as Quantile Data Structure [170] to ensure low latency for all computational operations and thus suitability for large data sets. Bernard et al. [25] proposed a system for cohort construction in temporal prostate cancer cohort data that included visualizations for subpopulations and individual patients. To guide users during exploration, visual markers indicate interesting relationships between attributes derived from statistical tests. Recently, Corvo et al. [56] presented a comprehensive visual analytics system for pathological high-throughput data, which encompasses all major steps of a typical data analysis pipeline, such as preprocessing raw histopathology images by interactive segmentation, components for exploratory data analysis, and interactive cohort construction in a high-dimensional feature space, feature engineering which includes extraction of potentially predictive biomarker features, modeling, as well as visualization and summarization of the modeling results. Preim and Lawonn provided comprehensive reviews of visual analytics methods and applications in public health [217] and epidemiology [219] in particular.

3.1.3 Previous Work on Subpopulation Discovery with the SHIP Data

Since we carry out the proof-of-concept for our workflow on the SHIP data, we list major scientific preparatory work hereafter. Preim et al. [218] provided an overview of the research that developed data mining and visual analytics methods to gain insights into the SHIP data. Among them is the "3D Regression Cube" of Klemm et al. [155], a system that allows interactive exploration of feature correlations in epidemiological datasets. The system generates many multiple linear regression models from different combinations of one dependent and up to three independent variables and displays their goodness of fit in a three-dimensional heat map. The system allows the user to modify the regression equation, for example, by changing the number of independent variables, specifying wild cards and interaction terms, fixing one of the variables to reduce computational complexity, or focussing specifically on a variable of interest. Our approach is also able to identify variables that are strongly associated with the target variable. However, we search for subpopulation-specific relationships rather than generating a global model for the entire dataset, and we additionally provide predictive value ranges. Klemm et al. [156] presented a system that combines visual representations of non-image and image data. They identify clusters of back pain patients for the SHIP data. Since we specify hepatic steatosis as the target variable, we instead build supervised models and classification rules that directly capture the relationships between predictors and the target variable. Alemzadeh et al. [4] presented S-ADVIsED, a system for interactive exploration of subspace clusters that incorporates various visualization types such as donut diagrams, correlation heatmaps, scatterplot matrices, mosaic diagrams, and error bar graphs. While S-ADVIsED requires the user to input mining results obtained in advance outside the system, our tool enables expert-driven interactive *subpopulation discovery* instead of expert-driven interactive *result exploration*. Hielscher et al. [126] 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 predictive feature subsets, the expert can provide their domain knowledge in the form of a small number of instance-level constraints, forcing pairs of instances (i.e., study participants) to be assigned either to the same or a different cluster. Hielscher et al. [125] extended their work and introduced a mechanism to validate subpopulations on independent cohorts.

Table 3.1: **Cost matrix.** Cost matrix to penalize misclassification under class imbalance.

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

3.2 Subpopulation Discovery Workflow and Interactive Mining Assistant

In this section, we present our subpopulation discovery workflow. We build classification models on the whole dataset and different partitions, as described in Section 3.2.1. In Section 3.2.2, we introduce relevant underpinnings of classification rule discovery, followed by a description of the primarily used HotSpot [110] algorithm in Section 3.2.3. We present our interactive mining assistant in Section 3.2.4. The dataset used for population partitioning and class separation on the target variable hepatic steatosis comes from the "Study of Health in Pomerania" (SHIP), recall Section 2.2.1. In Section 3.2.5, we describe the origin and availability of the target variable. In Section 3.2.6, the motivation for data partitioning and the partitioning steps are presented.

3.2.1 Classification

For the classification of cohort participants, we focus on algorithms that provide interpretable models, as we aim to identify predictive *conditions*, i.e., variables and values/ranges in the models. Therefore, we consider decision trees, classification rules, and regression trees. We use the J4.8 decision tree classification algorithm (equivalent to the C4.5 algorithm [220]) from the Waikato Environment for Knowledge Analysis (Weka) Workbench [82]. This algorithm builds a tree successively by partitioning each node (a subset of the dataset) to the variable that maximizes the information gain within that node. The original algorithm works only with variables that take categorical values, creating one child node per value. However, the Weka implementation also provides an option that forces the algorithm always to create exactly two child nodes: one for the best separating value and one for all other values. We use this option in our experiments because it yields better quality trees. The Weka algorithm also supports variables that take numeric values: A node is split into two child nodes by partitioning the variable's range of values into two intervals.

To deal with the skewed distribution, we consider 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 [47] 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 the following distribution A: 438, B: 216, C: 128.
- *CostMatrix*: We prefer to misclassify a negative case rather than not detecting a positive case, so we penalize false negatives (FN) more than false positives (FP). We use the cost matrix depicted in Table 3.1.

3.2.2 Classification Rule Discovery

Classification rules can reveal interesting relationships between one or more features and the target variable [87, 123]. Compared to model families such as deep neural networks, support vector machines and random forests, classification rules usually achieve lower accuracy. However, they are easier to interpret and infer and are therefore more suitable for interactive subpopulation discovery. In epidemiological research, interesting subpopulations could subsequently be used to formulate and validate a small set of hypotheses or investigate associations between risk factors for a particular target variable. A subpopulation of interest could be formulated as follows: "In the sample of this study, the prevalence of goiter is 32%, whereas the probability in the subpopulation described by *thyroid-stimulating hormone* less than or equal to 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 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:

$$r_1 : \underbrace{som\_waist\_s2 < 80 \wedge age\_ship\_s2 > 59\,(\wedge ...)}_{\text{Antecedent}} \longrightarrow \underbrace{hepatic\_steatosis = pos}_{\text{Consequent}} \tag{3.1}$$

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 side). 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 that balances coverage and confidence gain and is often used as an internal criterion for candidate generation [123]. It is defined as

$$WRA(r) = Cov(r) \cdot \left(Conf(r) - \frac{n_{T=v}}{N} \right). \tag{3.2}$$

The *odds ratio* of r with respect to $T = v$ is defined as

$$OR(r) = \frac{|s(r)_{T=v}|}{|s(r)_{T \neq v}|} \Big/ \frac{n_{T=v} - |s(r)_{T=v}|}{n_{T \neq v} - |s(r)_{T \neq v}|}. \tag{3.3}$$

As an example, Figure 3.3 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, $Sup(r_2) = 0.30$, $Conf(r_2) = 0.75$, $WRA(r_2) = 0.40 \cdot (0.75 - 0.40) = 0.14$ and $OR(r_2) = (3/1) \,/\, (1/5) = 15$.

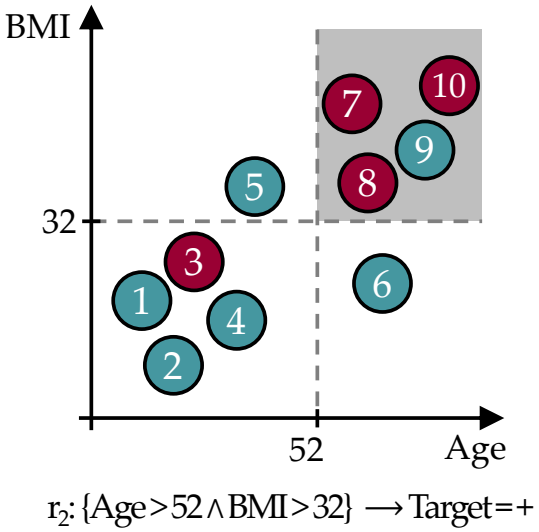


Figure 3.3: **Exemplary classification rule.** The gray area represents the data space of the covered instances.

3.2.3 HotSpot

For classification rule discovery, we use the HotSpot [110] algorithm provided for Weka [82]. HotSpot is a beamwidth search algorithm that implements a general-to-specific approach to rule extraction. A single rule is constructed by successively adding the condition to the antecedent that locally maximizes confidence. Unlike general hill-climbing, which considers only the best rule candidate at each iteration, HotSpot's beam search retains the b highest-ranked candidates and refines them in later steps. Consequently, HotSpot reduces the *"myopia"* [87] from which Hill-Climbing search typically suffers. Briefly, hill-climbing approaches consider only the locally optimal candidate at each iteration. As a result, a globally optimal rule will not be found if it is not locally optimal in each iteration. It is also desirable to generate more than one rule from an application perspective since alternative descriptions of subpopulations can facilitate hypothesis generation. The beamwidth can be specified as a `maximum branching factor`, i.e., the maximum number of conditions that can be added to a candidate rule. In each iteration, the rule candidates must satisfy the `minimum value count`, the sensitivity threshold. To avoid adding a condition only leads to a marginal improvement of the confidence, the parameter `minimum improvement`, i.e., the minimum relative improvement of the confidence by adding another condition, can be specified. The rule search's computational complexity can be reduced by specifying a `maximum rule length`, i.e., the number of conditions in the antecedent. In our experiments we set the parameters as follows: `maximum branching factor` = 20, `maximum value count` = 1/3, `minimum improvement` = 0.1, `maximum rule length` = 3.

3.2.4 Interactive Medical Miner

Classification rules can provide valuable insights into potentially prevalent conditions for different subpopulations of the cohort under study. However, when the number of rules created is large, as is usually the case with large epidemiological data, the rules' conditions overlap. Hence, some conditions are present under each of the classes of the target variable. Therefore, the medical expert needs inspection tools to decide which rules are informative and which features should be investigated further. Our Interactive Medical Miner (IMM) allows the expert

to

- discover classification rules,
- inspect the frequency of these rules (a) against each class and (b) against the unlabeled subset of the cohort, and
- examine the statistics of each rule for the values of selected variables.

We describe these functionalities below, referring to the screenshot in Figure 3.4.

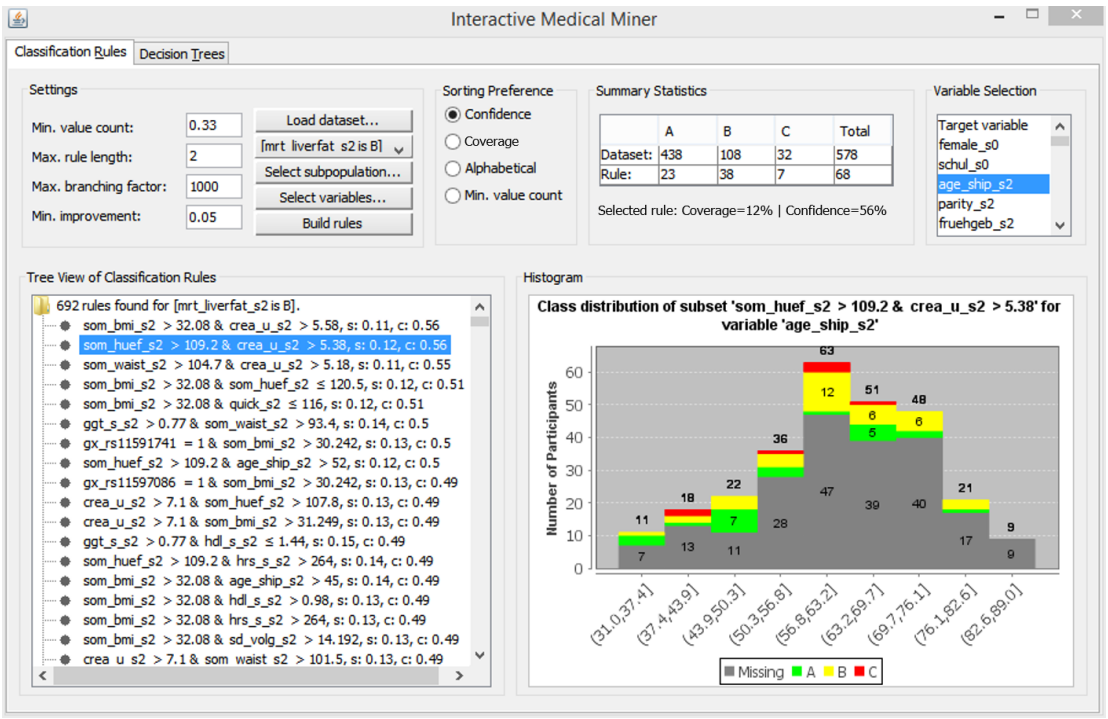


Figure 3.4: **The user interface of the Interactive Medical Miner.** Classification rules are discovered for class B and shown in the bottom left panel. For the selected rule som_huef_s2 > 109 & crea_u_s2 > 5.38 ⟶ mrt_liverfat_s2 = B, the distribution of the participants covered by the rule among all three classes is shown in absolute values (top middle panel) and as a histogram (bottom right panel) with respect to age (top right panel).

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

Before rule generation, the user can specify a sub-cohort of the dataset. By clicking on the button `Select Subpopulation`, a popup window appears, where multiple filter queries in the form of `<variable> <operator> <value>` can be added, e.g. `som_bmi_s2 >= 30`. The defined constraints are displayed in a table and can be undone. Furthermore, the user can select variables for model creation, e.g., exclude a variable that is already known to be highly correlated with another variable that is already considered for model learning.

Mining criteria include the dataset (choose between the whole dataset and one of the partitions), the class for which rules are to be generated (drop-down list "Class"), and the constraints related

to this class, i.e., "Minimum number of values" (which can also be specified as a relative value), "Maximum rule length", "Maximum branching factor" and "Minimum improvement". As an example of how these parameters affect rule search, consider the selected rule in Figure 3.4, som_huef_s2 > 109 & crea_u_s2 > 5.38 ⟶ 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 evident that each of the two conditions in the antecedent of the rule, namely som_huef_s2 > 109 and crea_u_s2 > 5.38, must also exceed this threshold. The position of a condition within the antecedent indicates at which refinement step the condition was added to the rule candidate. For example, the first condition som_huef_s2 > 109 with a confidence of 44/107 = 0.41 was extended by the second condition crea_u_s2 > 5.38 because the confidence gain exceeds the minimum improvement threshold, i.e., 38/68 - 44/107 = 0.15 > 0.05. However, this rule cannot be extended further because the maximum rule length is set to 2. The maximum branching factor was conservatively set to 1000 to prevent potentially interesting rules from not being generated due to a small beamwidth. The expert can lower this parameter interactively if the number of rules found is too 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 upper-middle area "Summary Statistics" is updated. The first row shows the distribution of cohort participants across classes for the entire dataset. The second row shows how the participants covered by the rule (column "Total" in the second row) are distributed across classes. Thus, the expert can specify the discovery of classification rules for one of the classes and then examine how often each rule's antecedent occurs among participants in the other classes. For example, a rule that covers most of the participants in the selected class (class B in Figure 3.4) is not necessarily interesting if it also covers a high number of participants in the other classes. The rule som_huef_s2 > 109 & crea_u_s2 > 5.38 ⟶ mrt_liverfat_s2 = B covers a total of 68 participants, of which 38 are of class B. To reduce the number of covered participants from other classes, i.e., to increase the confidence, the user can decrease the minimum value count threshold to allow generating rules with a lower sensitivity but more homogeneity with respect to the selected class.

Some of the data may be incomplete. For example, not all participants in the cohort underwent liver MRI. Therefore, it is also of interest to know the distribution of unlabeled participants who support a given rule's antecedent. For this purpose, the "Histogram" panel can be used: The expert selects another feature from the interactive "Variable selection" area in the upper right panel and can then see how the values of this variable are distributed among the study participants – both labeled and unlabeled; the latter are marked as "Missing" in the color legend. For plotting the histograms, we use the free Java chart library JFreeChart [94]. Numerical variables are discretized using "Scott's rule" [243] 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 $h(X_{s(r)}) = \frac{\max X_{s(r)} - \min X_{s(r)}}{3.49\sigma_{s(r)}} \cdot |s(r)|^{\frac{1}{3}}$.

If the expert does not select a variable, the target variable is used by default, and only the distribution of labeled participants is visible. The histogram in Figure 3.4 shows the age distribution of both labeled and unlabeled participants covered by our example rule som_huef_s2 > 109 & crea_u_s2 > 5.38 ⟶ mrt_liverfat_s2 = B. The distribution of values among the labeled participants indicates that age may be a risk factor for the indicated subpopulation, as the probability of class B increases with age. This visual finding suggests adding the condition age_ship_s2 > 56.8 to the antecedent of the rule. Indeed, the confidence of this more specific rule increases from 38/68 = 0.56 to 27/40 = 0.675. However, as the sensitivity decreases from 38/108 = 0.352 to 27/108 = 0.250, the minimum value count threshold is no longer met. Thus, visualizing participant statistics for selected rules can provide clues to subpopulations that should be monitored more closely and clues to how to modify algorithm parameters for

subsequent runs, in our example, to decrease the minimum value count to 0.25 and increase the maximum rule length to 3.

3.2.5 The Target Variable

The target variable is derived from participants' liver fat concentration calculated by magnetic resonance imaging (MRI). At the time of writing the original manuscript, MRI results were only available for 578 (from a total of 2333; ca. 24.7%) SHIP-2 participants. We use the data from these participants for classifier learning, while our Interactive Medical Miner also contrasts these data with data from the remaining 1755 participants for whom MRI scans were not made available.

After discussions with domain experts, we decided to assign participants with a liver fat concentration of 10% or less to class A ("negative" class, i.e., absence of the disorder); values greater than 10% and less than 25% represent class B (increased liver fat/fatty liver tendency) and values greater than 25% class C (high liver fat). We consider classes B and C as "positive". The cutoff value of 10% is intentionally higher than the value of 5% proposed by Kühn et al. [164] to separate subjects with and without hepatic steatosis because the primary interest from a medical perspective was to identify predictive variables for subjects likely to be ill. Selecting a high cutoff value exacerbates class imbalance and makes data analysis more difficult. Figure 3.5 depicts the class distribution stratified by sex. Of the 578 participants, 438 belong to class A (approximately 76%), 108 to B (19%), and 32 to C (6%). Men were more likely to have elevated or high liver fat concentration than women (30.7% vs. 18.8% in classes B or C).

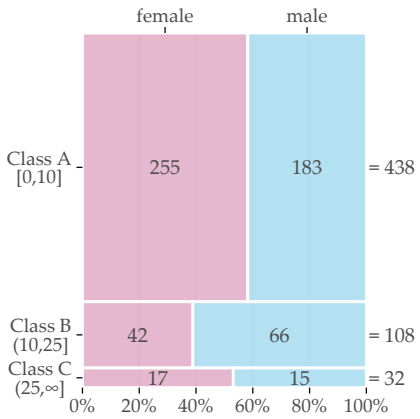


Figure 3.5: **Sex-specific distribution of the target variable.** The boxes' relative sizes depict the number of female and male participants for each of the classes.

In addition to the target variable, the data set contains 66 variables extracted from participants' questionnaire responses and medical tests (cf. [282]). These are variables on sociodemographics (e.g., sex and age), self-reported lifestyle indicators (e.g., alcohol and cigarette consumption), single-nucleotide polymorphism (SNP), laboratory measurements (e.g., serum concentrations), and liver ultrasound. The two available variables of liver ultrasound are `stea_s2` and `stea_alt75_s2`. Both take symbolic values reflecting the probability that the participant has a fatty liver; the latter is a combination of the former and the participant's alanine transaminase (ALAT) concentration; details are in the caption of Table 3.3 and in [282]. Almost all variables mentioned below have the suffix `_s2` indicating SHIP-2 follow-up measurements, in contrast to SHIP-0 (`_s0`) and SHIP-1 (`_s1`). Exceptions are sex, highest school degree, and the 10 SNP variables.

3.2.6 Partitioning the Dataset into Subpopulations

Because the dataset is imbalanced with respect to sex (314 females, 264 males), we decided to partition the dataset before classification. First, we examined the class distributions for each sex. We observed that the distributions were very different, especially for class B (see Figure 3.5). Second, we examined the class distribution by sex and age. We found that age was associated with the female subpopulation, but not with the male subpopulation. Third, we identified a cutoff point for age by introducing a heuristic that determines the age value that minimizes the target variable's standard deviation. We then performed supervised learning separately on the partitions of female and male participants, referred to as `PartitionF` and `PartitionM` hereafter. We also created an additional learner for the subpopulation of older female participants aged above the cutoff point of 52 (Partition `F:age>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 $X_y = \{x \in \texttt{PartitionF} | \text{age of } x \leq splitAge\}$, $X_z = \{x \in \texttt{PartitionF} | \text{age of } x > splitAge\}$ 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

$$SwSD\left(X_y, X_z\right) = \frac{|X_y|}{n}\sigma(X_y) + \frac{|X_z|}{n}\sigma(X_z) \tag{3.4}$$

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.

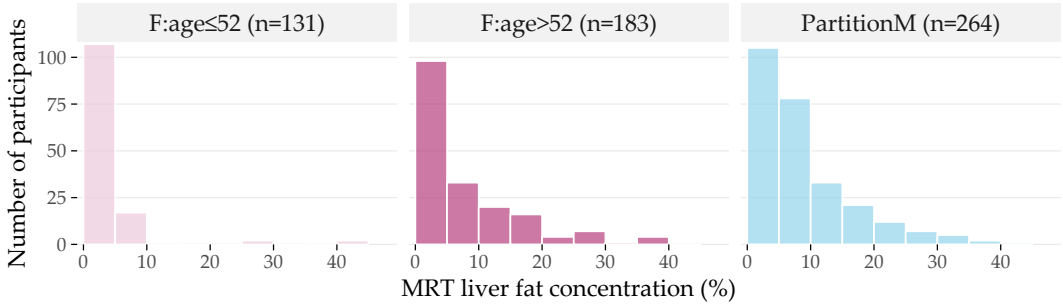


Figure 3.6: **Distribution of liver fat concentration for each partition.** Distribution of liver fat concentration in male participants (`PartitionM`), and females younger and older than 52 years. The horizontal axis shows the liver fat concentration in bins of 5%, while the vertical axis indicates the number of participants in each bin.

The histograms in Figure 3.6 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 `F:age>52` (n=183) of `PartitionF`. Most of the 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 `F:age>52` have a liver fat concentration of more than 10%; they belong to the positive classes B and C.

3.3 Experiments and Findings

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 (%)	F-measure (%)
`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

3.3.1 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-measure, i.e., the harmonic mean between precision and recall. We consider the two classes B and C together as the positive class for specificity, precision, and recall.

Oversampling achieved the best performance with an accuracy of about 80% and an F-measure score of 62%. We found the best decision trees for `F:age>52`, followed by those for `PartitionF`, then `PartitionM`. The large discrepancy between the accuracy and F-measure scores also appears in the partitions' models, suggesting that the accuracy scores are unreliable in such a skewed distribution. Therefore, we do not report on accuracy below.

On partition `F:age>52`, the overall best decision tree is achieved by the oversampling variant. On the larger `PartitionF`, the best performance was achieved by the decision tree created with the InfoGain variant. In contrast, the best decision tree on `PartitionM` was created with the CostMatrix variant. The sensitivity and specificity values for these trees are given in Table 3.2, while the trees themselves are shown in Figures 3.7 - 3.9 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 for the class imbalance problem. As `PartitionM` has the most heterogeneous class distribution out of all partitions, all variants perform relatively poorly 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 to report 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 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.

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 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

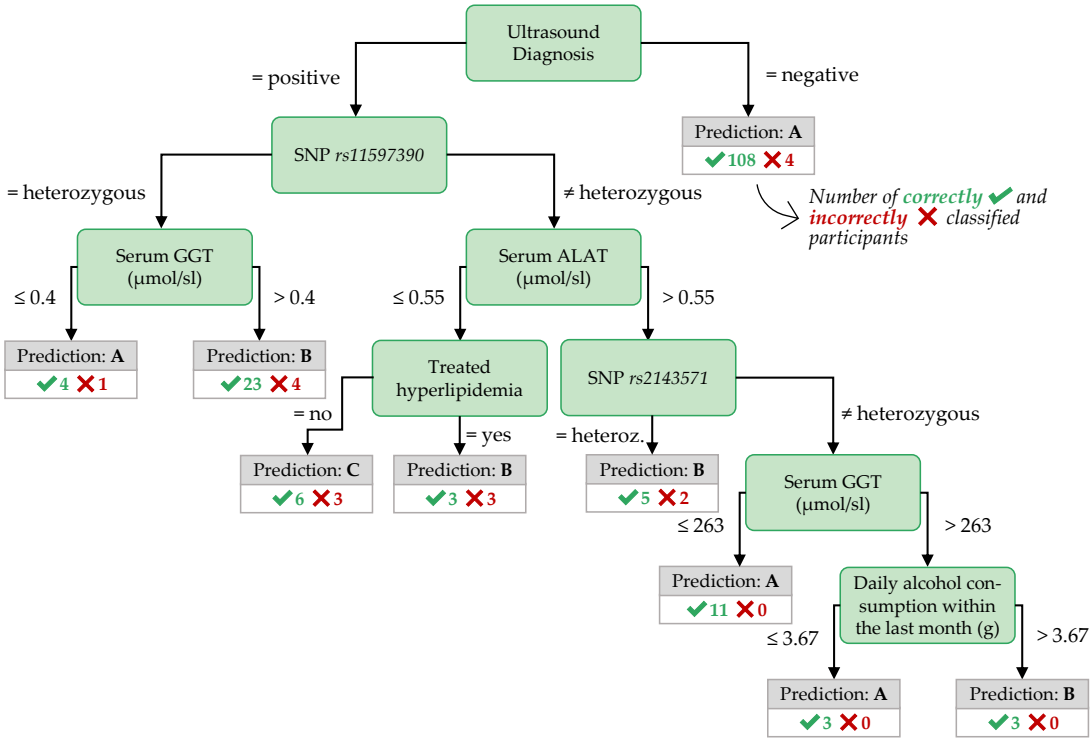


Figure 3.7: **Best decision tree for `F:age>52`**, achieved by the variant *Oversampling*.

stea_s2. A negative ultrasound diagnosis points to negative class A, but a positive ultrasound diagnosis does not directly lead to the positive classes B and C. The decision trees of the three partitions differ in the nodes placed near the root.

Important features for PartitionF. In the best decision tree of `PartitionF` (cf. Figure 3.8), it can be observed 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.

Important features for F:age>52 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 shown in Table 3.4. Similar to `PartitionF`, high BMI values point to a positive class when combined with other features. Table 3.4 shows 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 with a positive ultrasound

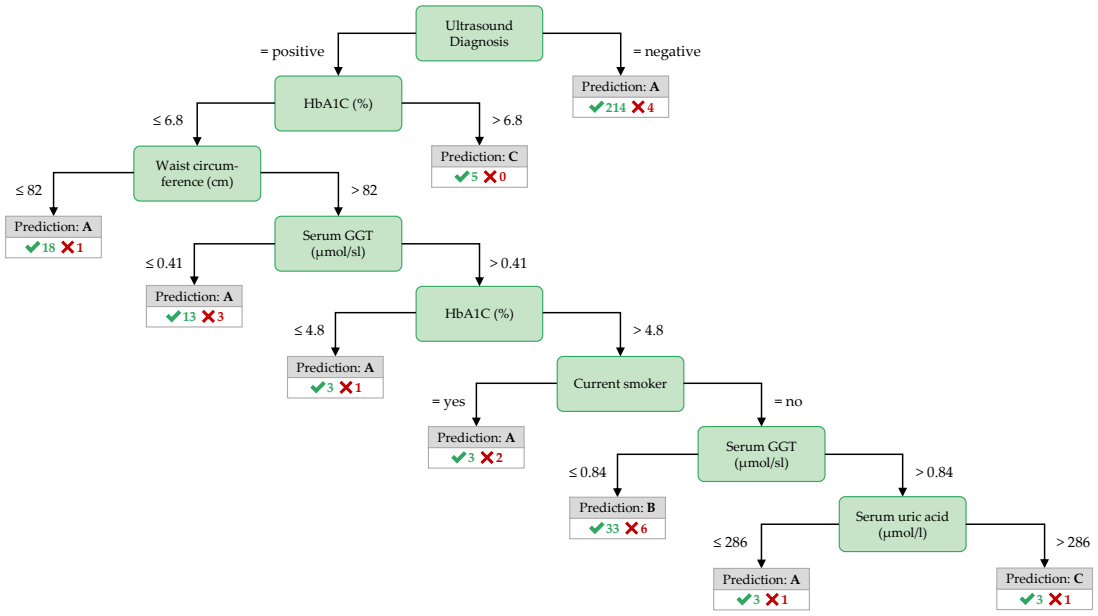


Figure 3.8: **Best decision tree for `PartitionF`**, achieved by the variant *InfoGain*.

diagnosis, a genetic marker gx_rs11597390 = 1, and a high-density lipoprotein (HDL) serum concentration of at most 1.53 mmol/l.

Important features for PartitionM. 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 Gamma-glutamyltransferase (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 mol/sl in combination with creatinine concentration of at most 90 mmol/l or a thromboplastin time ratio (quick_s2) of more than 59% points 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 provide insights into features that appear diagnostically important. However, the medical expert needs additional information to decide whether a feature is worth further investigation. Decision trees highlight the importance of a feature only in the context of the subtree in which it is found; a subtree describes a typically very small subpopulation. In contrast, classification rules provide information about 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.

Furthermore, unless a classification rule has a confidence value close to 100%, participants in the other classes may also support it. Therefore, to decide whether the features in the antecedent of the rule deserve further investigation, the expert also needs knowledge about the statistics of the rule for the other classes. To assist the expert in this task, we have proposed the Interactive Medical Miner. This tool discovers classification rules for each class *and* provides information about the statistics of these rules for all classes.

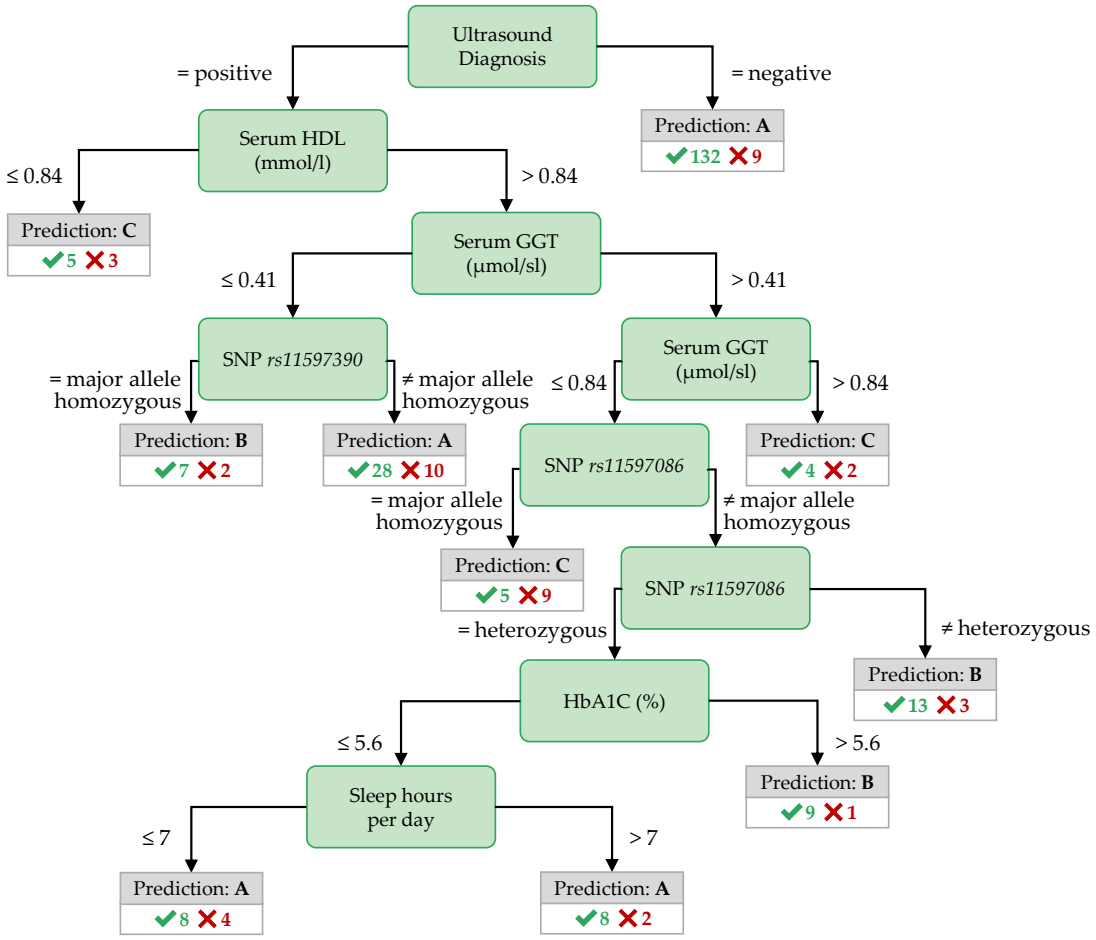


Figure 3.9: **Best decision tree for `PartitionM`**, achieved by the variant *CostMatrix*.

3.4 Conclusion

To date, analysis of population-based cohort data has been mostly hypothesis-driven. We have presented a workflow and an interactive application for data-driven analysis of population-based cohort data using hepatic steatosis as an example. Our mining workflow includes steps

i. to discover subpopulations that have different distributions with respect to the target variable,
ii. to classify each subpopulation taking class imbalance into account, and
iii. to identify variables associated with the target variable.

Our workflow has shown that it is appropriate (a) to build subpopulations before classification to reduce class imbalance and (b) to drill-down on the derived models to identify important variables and subpopulations worthy of further investigation.

To assist the domain expert with the latter objective (b), we have developed the Interactive Medical Miner, an interactive application that allows the user to explore classification rules further and understand how the cohort participants supporting each rule are distributed across

Table 3.3: **Classification rules (`PartitionF`).** Best HotSpot classification rules (*maxLength* = 3) for `PartitionF` (excerpt). Cov: coverage; Sup: support; Conf: confidence. age_ship_s2: age; blt_beg_s2: time of blood sampling; ggt_s_s2: serum Gamma-glutamyltransferase (GGT; mol/sl); gluc_s_s2: serum glucose (mmol/l); gx_rs11597390: genetic marker; hrs_s_s2: serum uric acid concentration (µmol/l); ldl_s_s2: serum low-density lipoprotein (LDL; mmol/l); sleeph_s2: sleep hours; sleepp_s2: sleep problems; som_bmi_s0: body mass index; som_huef_s0: hip circumference (cm); som_waist_s2: waist circumference (cm); stea_alt75_s2: hepatic steatosis (ultrasound diagnosis) and alanin aminotransferase (ALAT) concentration 0.55 µmol/sl – 0 = normal, 1 = hypoechogenic, 2 = hyperechogenic, 3 = questionable; stea_s2: hepatic steatosis (ultrasound diagnosis); tg_s_s2: serum triglycerides (mmol/l); tsh_s2: thyroid-stimulating hormone (TSH; mu/l).

Rule antecedent			**Cov**	**Sup**		**Conf**
Variable 1	**Variable 2**	**Variable 3**	**Abs**	**Abs**	**Rel (%)**	**Rel (%)**
Target class: A						
som_waist_s2 ≤ 80	–	–	132	132	52	100
som_bmi_s2 ≤ 24.82	–	–	109	109	43	100
som_huef_s2 ≤ 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 ≤ 66	asat_s_s2 > 0.22	6	6	35	100
som_bmi_s2 > 38.42	sleeph_s2 > 6	blt_beg_s2 ≤ 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

the three classes. This exploration step is essential for identifying not-yet-known associations between some variables and the target. These variables must then be further investigated – in hypothesis-driven studies. Therefore, our workflow and Interactive Medical Miner carry the potential of data-driven analysis to provide insights into a multifactorial disease and generate hypotheses for hypothesis-driven studies. Our Interactive Medical Miner has been extended by Schleicher et al. [238], who added panels that include tables showing additional rule statistics such as lift and p-value. Besides, a mosaic plot contrasts the class distributions of a subpopulation and its "complements," i.e., subsets of participants who do not meet one or both conditions of a length-2 rule describing that subpopulation.

In terms of the multifactorial disorder of interest, our results confirm the potential of our data-driven approach because most of the variables in the top positions of our decision trees and classification rules have been previously shown to be associated with hepatic steatosis in independent studies. In particular, indices of fat accumulation in the body (BMI, waist circumference) and the liver enzyme GGT were proposed by Bedogni et al. [21] as a reliable "Fatty Liver Index". According to Yuan et al. [300], the SNPs rs11597390, rs2143571, and rs11597086 are among the "Independent SNPs Associated with Liver-Enzyme Levels with Genome-wide Significance in Combined GWAS Analysis of Discovery and Replication Data Sets". Regarding

Table 3.4: **Classification rules `F:age>52`.** Best HotSpot classification rules (*maxLength* = 3) for `F:age>52` (excerpt). Cov: coverage; Sup: support; Conf: confidence. age_ship_s2: age; crea_u_s2: urine creatinine (mmol/l); fib_cl_s2: fibrinogen (Clauss) (g/l); gluc_s_s2: serum glucose (mmol/l); ggt_s_s2: serum Gamma-glutamyltransferase (GGT; mol/sl); gx_rs11597390: genetic marker; hdl_s_s2: high-density lipoprotein (mmol/l); hrs_s_s2: serum uric acid concentration (µmol/l); som_bmi_s0: body mass index; som_huef_s0: hip circumference (cm); som_waist_s2: waist circumference (cm); stea_alt75_s2: hepatic steatosis (ultrasound diagnosis) and alanin aminotransferase (ALAT) concentration 0.55 µmol/sl – 0 = normal, 1 = hypoechogenic, 2 = hyperechogenic, 3 = questionable; stea_s2: hepatic steatosis (ultrasound diagnosis).

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

the effects of alcohol consumption, toxic effects of alcohol on the liver are well established and ascribe an even more significant role to obesity than heavy alcohol consumption concerning fat accumulation in the liver [20, 22]. Indeed, a variable related to alcohol consumption appears only in our decision tree on `F:age>52` (see Figure 3.7) and not among our top classification rules, where we tend to see variables associated with a person's weight and obesity (cf. variables som_bmi_s2, som_huef_s2, som_waist_s2 in all figures and tables in Section 3.3). The subpopulation `F:age>52` itself was identified without prior knowledge of this subpopulation's semantics. Still, it is noteworthy that the age of 52 years is close to the onset of menopause – Völzke et al. [285] showed that menopausal status is associated with hepatic steatosis. Our results also verify another fact that was known to medical experts by independent observation: the sonographic variables (cf. stea_s2, stea_alt75_s2 in all figures and tables in Section 3.3) is associated with liver fat concentration found on MRI, but ultrasound alone does not predict hepatic steatosis [22, 21].

Our algorithms not only provide variables associated with the target but also identify the value intervals related to a specific class, see, for example, the value intervals of BMI associated with class B for `PartitionM` (Table 3.5) and with classes A and C for `F:age>52` (Table 3.4). These intervals do not imply that a person with a BMI within the specific interval actually belongs to the corresponding class but can serve as a starting point for hypothesis-driven analyses.

Table 3.5: **Classification rules (`PartitionM`).** Best HotSpot classification rules (*maxLength* = 3) for `PartitionM` (excerpt). Cov: coverage; Sup: support; Conf: confidence. age_ship_s2: age; ATC_C09AA02_s2: enalapril intake; chol_s_s2: serum cholesterol (mmol/l); crea_u_s2: urine creatinine (mmol/l); crea_s_s2: serum creatinine (µmol/l); fig_cl_s2: Fibrinogen (Clauss) (g/l); ggt_s_s2: serum Gamma-glutamyltransferase (GGT; mol/sl); gout_s2: treated gout (self-report); hdl_s_s2: high-density lipoprotein (mmol/l); hgb_s2: haemoglobin (g/l); hrs_s_s2: serum uric acid concentration (µmol/l); jodid_u_s2: urine iodide (µg/dl); quick_s2: thromboplastin time Quick test (%); sleeph_s2: sleep hours; stea_alt75_s2: hepatic steatosis (ultrasound diagnosis) and alanin aminotransferase (ALAT) concentration 0.55 µmol/sl – 0 = normal, 1 = hypoechogenic, 2 = hyperechogenic, 3 = questionable; stea_s2: hepatic steatosis (ultrasound diagnosis); som_bmi_s0: body mass index; som_huef_s0: hip circumference (cm); som_waist_s2: waist circumference (cm); tg_s_s2: serum triglycerides (mmol/l).

	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

Our approach allows us to study subpopulations at two points in time. Before the modeling step, we identify subpopulations that have different class distributions. During the modeling step, our Interactive Medical Miner highlights the subpopulation that supports each classification rule; these are *overlapping* subpopulations. Overlapping subpopulations are not necessarily a disadvantage, especially for very small subpopulations. However, working with overlapping data sets can be unintuitive and tedious for a domain expert. In Chapter 4, we explore the potential of clustering to identify and reorganize *overlapping* rules to reduce the user's cognitive load by displaying only semantically unique and representative rules.

Although the workflow's main application is a *longitudinal* cohort study, it does not exploit temporal characteristics in the data. However, indicators of lifestyle change are potentially predictive for the later occurrence of a disease. Chapter 6 presents a follow-up method that expands the feature space by extracting temporal variables that describe how a study participant changes over time to derive new informative variables for hypothesis generation.

Chapter 4

Identifying Distinct Subpopulations

Brief Chapter Summary

Subgroup discovery algorithms often generate redundant descriptions of subpopulations. We present a workflow to extract a small number of representative rules from a large set of classification rules. These so-called *proxy rules* minimize instance overlaps across rule groups, thus covering different subpopulations. We evaluate our workflow on SHIP data samples with hepatic steatosis and goiter as target variables, respectively.

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: *Computer-Based Medical Systems (CBMS)*. 2017, pp. 582-587. DOI: 10.1109/CBMS.2017.15.

Epidemiologists search for significant relationships between risk factors and a target variable in large and heterogeneous datasets that encompass participant health information gathered from questionnaires and medical examinations [69]. The previous chapter describes an expert-driven workflow that can help epidemiologists automatically detect such relationships in the form of classification rules, which are descriptions of risk factors and predictive value ranges for a specific subpopulation and an target variable of interest. However, rule induction algorithms often produce large and overlapping rule sets requiring the expert to manually pick the most informative rules and remove less informative or redundant ones. 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 essential subpopulations while maintaining *distinctiveness* between the clusters. For each cluster, a representative rule is shown to the expert who then can drill down to other cluster members. We evaluate our algorithm on two subsets of SHIP where the target variables hepatic steatosis and goiter serve as target variables, respectively. Further, we report on the effectiveness of our algorithm, and we present selected subpopulations.

We propose SD-Clu, an approach that combines subgroup discovery with clustering to return k representative classification rules. Building upon a set of potentially highly overlapping rules generated by an 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 tradeoff between rule confidence and coverage towards the target variable. We define the 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.

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 describes the experimental setup. In Section 4.4, we reports on our evaluation results. In Section 4.5, we discuss our findings regarding hepatic steatosis and goiter on the SHIP data. In Section 4.6, 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 [123, 11]. Compared to more accurate but predominantly opaque black-box models such as neural networks, support vector machines, or random forests, SD algorithms yield more interpretable 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 [198], to identify predictive auditory-perceptual, speech-acoustic, and articulatory-kinematic features of preschool children with speech sound disorders [278], and to discover discriminative features in patient subpopulations with different admission times to psychiatric emergency departments [41].

However, SD methods often yield large sets of rules that domain experts are not willing to tediously go through and manually separate interesting from irrelevant or redundant ones. 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 can explore a compact set of different subpopulations, equipped with mechanisms to drill down to specific rules of interest.

Similar to the HotSpot algorithm described in Section 3.2.3, popular SD algorithms such as SubgroupMiner [157], SD [88], and CN2-SD [175] use a fixed beamwidth [87] to limit the number of further expanded subgroup candidates at each iteration. A post-pruning step can be applied to reduce the size of the rule set – e.g., to return the *top k* rules – using a quality criterion such as the Weighted Relative Accuracy [174] or the p-value of a statistical test. Even when both beamwidth 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 subpopulation (i.e., high instance overlap among rules) [176].

Unlike SD algorithms which generate multiple rules that overlap in terms of their coverage sets, predictive rule learning algorithms such as CN2 [51] or RIPPER [54] are designed to generate rules that capture different subpopulations. They work iteratively according to a divide-and-conquer strategy [87] as follows: First, a rule that maximizes the algorithm's quality function is

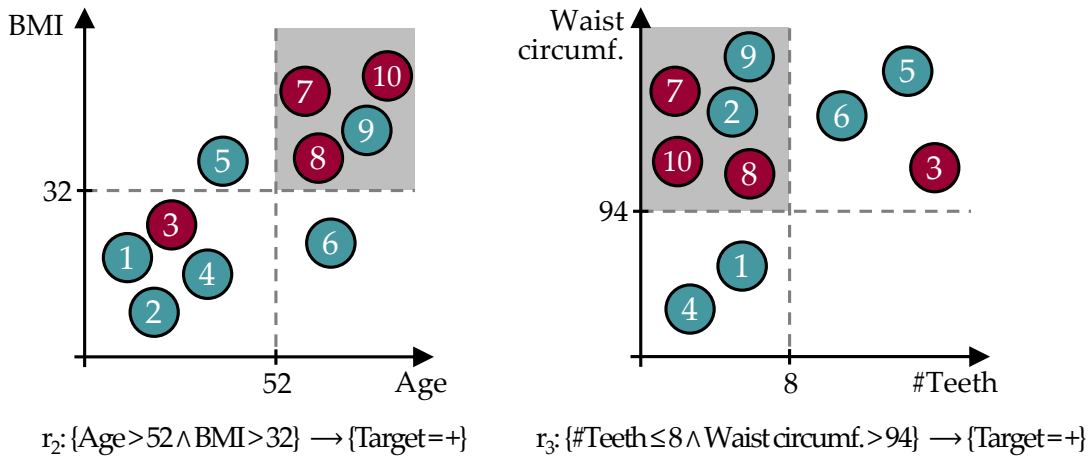


Figure 4.1: **Example of *redundant* rules.** Both r_2 and r_3 cover the instances 7,8,9, and 10; r_3 additionally covers instance 2. The cover set overlap is due to the high correlation between *#Teeth* and *Age*; and between *BMI* and *Waist circumference*. Both rules describe the same subpopulation, i.e., *elderly overweight people are at higher risk of developing the disease*.

generated on the set of instances not yet covered. Second, all covered instances are removed from the training 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 rule redundancy, important subpopulations may remain undiscovered because of order dependencies. For example, the algorithm might induce a rule that includes instances with a high BMI, but it might not find a slightly weaker association with income because those instances are immediately removed from the instance candidate space. Also, 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 merely eliminating covered instances from rule induction of subsequent iterations, *weighted covering approaches* consider how many times instances have been covered so far in the rule candidate expansion step [175]. While this leniency in 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. It may not make sense to interpret a single rule.

4.2 Finding Distinct Classification Rules

This section presents 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 distinct rule clusters covering different sets of instances. For each cluster, the rule with the best tradeoff between confidence and coverage of the target variable is appointed as the representative of the group. Rules covering

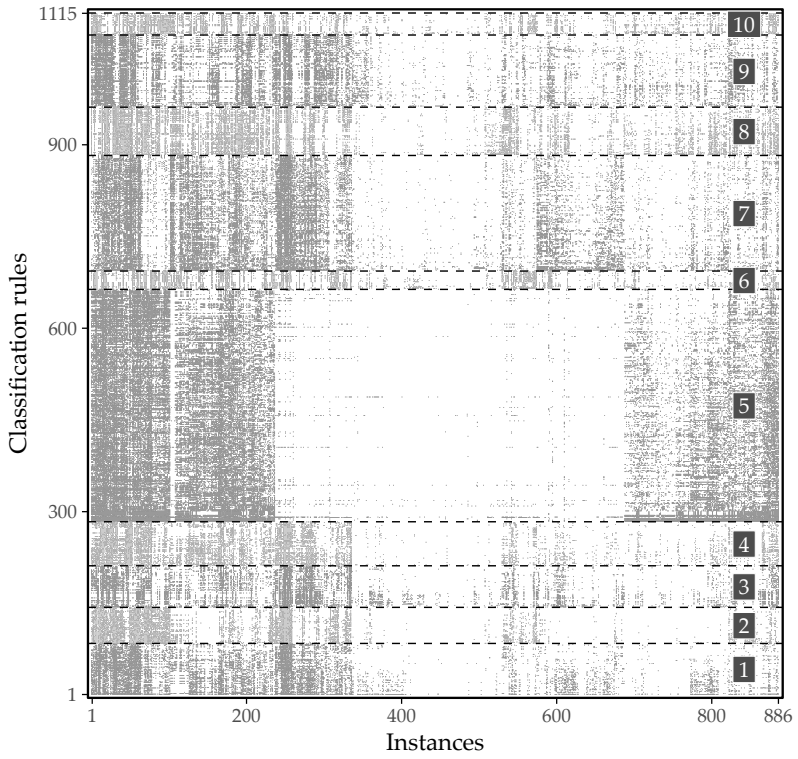


Figure 4.2: **Illustration of rule redundancy.** Graphical representation of 1115 HotSpot rules found on SHIP data on 886 labeled instances. A gray cell indicates that the instance at that x-position is covered by the rule at the associated y-position. Instances and rules are sorted by agglomerative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

the same instances are grouped in the same cluster and are therefore represented by the same proxy rule.

4.2.1 Rule Clustering and the Concept of "Proxy Rules"

We use the same notation for classification rule discovery as in Section 3.2.1. Agglomerative hierarchical clustering iteratively merges clusters of similar instances in a bottom-up way. Here, the instances are rules. Hence an alternative definition of a distance measure is required. 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 is selected for merging.

We define rule similarity for clustering on the basis of the mutually covered instances as an adaption of the Sørensen–DICE coefficient [63]. The distance between two rules r_1, r_2 with corresponding subpopulations (cover sets) $s(r_1)$, $s(r_2)$ is given as

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

We propose two ways of completing the hierarchical rule clustering process:

1. Specification of the maximum number of clusters k
2. Discovery of the optimal k based on an internal clustering index such as the Silhouette coefficient.

For a clustering ξ and a set of rules R, the Silhouette coefficient is calculated as

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

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. Basically, any other cluster quality function could be used. We choose the Silhouette coefficient here because of its understandable interpretation. The value of the silhouette coefficient can vary between -1 and 1. A negative value is undesirable because it represents a case where the average distance to rules in the same cluster is greater than the minimum average distance to rules in another cluster. We prefer a positive silhouette coefficient, i.e., a value close to the maximum of 1.

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 [174] (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 Proxy Rules

The proxy rules 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 the 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 proxy rules' representativeness 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 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 a data 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 $bin_0(U) = \{x \in X | \forall r \in U : isCovered(x, r) = 0\}$. 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. Therefore, we 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 proxy rules R_ζ, the *representativeness* of R_ζ is defined as

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

We investigate how the *representativeness* changes with an increasing number of proxy rules k for different criteria to select the cluster's proxy rule. Ideally, *representativeness* is high while k is low. In other words, the number of proxy rules should be as small as possible, but these rules should cover all important subpopulations within the data. For a fixed k, we compare our strategy of selecting the cluster's rule with the highest WRA as the proxy rule against three baselines: the top k rules according to (i) WRA, (ii) odds ratio, and (iii) coverage.

4.3 Experimental Setup

Datasets. To evaluate our method, we used data from the SHIP study. For a description of SHIP, see Section 2.2.1. We considered hepatic steatosis (see Section 3.2.5) and goiter as target variables. For the sample `HepStea`, we derived a binary target variable by discretizing the liver fat concentration obtained from the MRI report, such 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 expressed ten years later in SHIP-2. For the `Goiter` sample, the target variable was derived by thyroid ultrasound. The presence of goiter was defined for a thyroid volume greater than 18 ml in women and 25 ml in men [107]. Of the 4400 participants for whom the target 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, we use our approach from Chapter 3, the algorithm HotSpot [111] (Section 3.2.3). We further try SD-Map [13], an exhaustive algorithm that adapts the popular FP-Growth association rule learning method [113]. Rules that fall below a minimum coverage threshold are pruned. A depth-first search is performed for candidate generation. Rules are ranked according to a user-defined quality function. We use the implementation from

Table 4.1: **Statistics of best runs per dataset and algorithm.** Number of rules $|R|$, optimal Silhouette coefficient $Silh(\zeta_{opt})$, the corresponding number of rule proxies k of the optimal clustering $|\zeta_{opt}|$ and percentage of rule 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	21.3
Goiter	SD-Map	106	0.66	54	50.9

the VIKAMINE framework [12]. The implementation of SD-Map only supports categorical variables. Therefore, each numeric variable is discretized using the minimum description length based approach of Fayyad and Irani [75].

For SD-Map, we set the minimum coverage threshold to 0.05 to avoid overfitting, too small rules. 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 beamwidth is set to 500. Furthermore, 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. To determine the optimal number of proxy rules k, we compute the Silhouette coefficient for every possible number of clusters.

4.4 Results

Figure 4.3 shows the optimal number of clusters for each study sample combination and SD algorithm. Table 4.1 lists the optimal k and the ratio of proxy rules vs. the total number of rules. For example, clustering with optimal Silhouette coefficient for the algorithm HotSpot on `Goiter` has 76 clusters and thus 76 proxy rules (cf. Table 4.1), which is only 21.3% of the total number of rules. Thus, if only the cluster proxies are initially displayed to the expert, the time needed to check the rules is reduced by 78.7%.

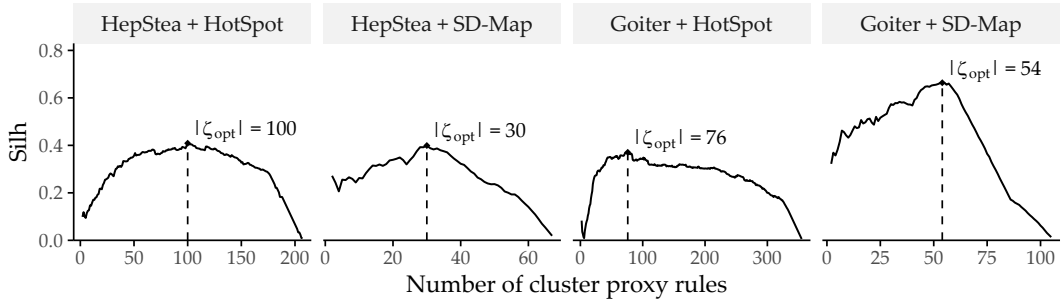


Figure 4.3: **Silhouette coefficients ($Silh$) of SD-Clu using complete linkage for each combination of dataset and algorithm.** The number of clusters k with the highest $Silh$ score ($|\zeta_{opt}|$) is indicated by a dashed vertical line.

The optimal number of clusters k 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 wish to obtain. For example, if the

expert considers $|\zeta_{opt}| = 100$ too large for HotSpot on `HepStea`, the diagram would show them 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 relatively 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 plot matrix' panels are arranged by rule selection method (rows), and the number of representative rules k returned to the expert (columns). Each graph shows the $adjCovRate$ (y-axis) of instances (x-axis) for 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) [52] of the points. Ideally, both curves are close to each other, meaning that the instances are covered by cluster proxy rules approximately by the same proportion as all rules cover them.

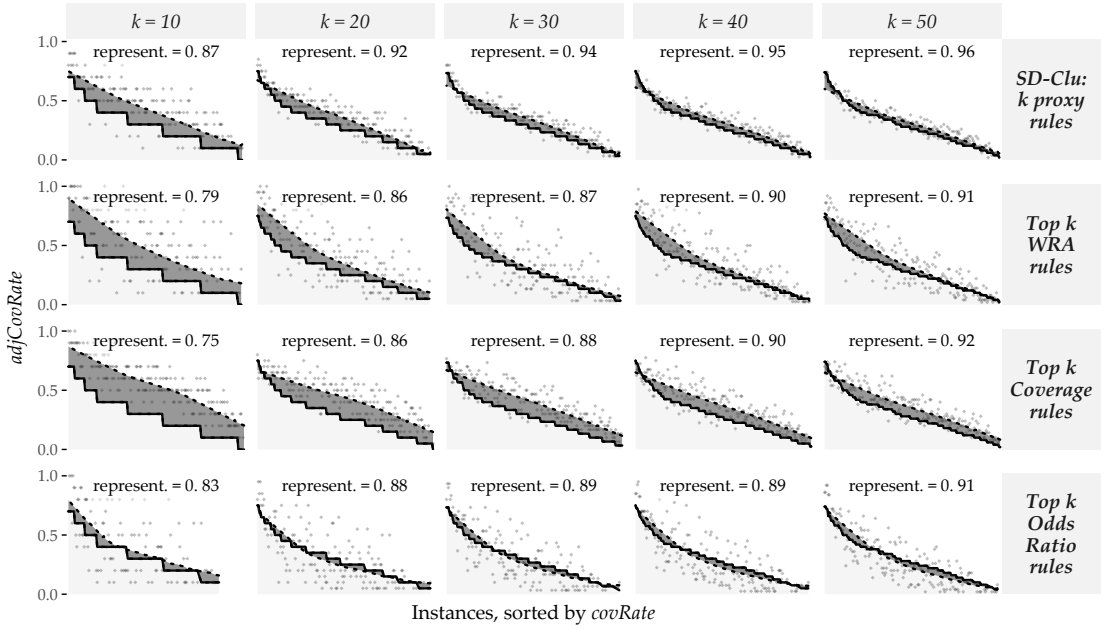


Figure 4.4: **Evaluation of *representativeness* on `HepStea` using the HotSpot algorithm.** *representativeness* of SD-Clu and three baseline approaches for different numbers of clusters k on the `HepStea` sample using HotSpot. Points depict an instance's $adjCovRate$ for the set of representative rules 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 curves illustrates *representativeness* of the top k rules of the respective approach, illustrated by the dark gray area in-between solid curve and dotted curves. Smaller areas are better and reflect higher *representativeness* values.

For all approaches, *representativeness* improves as k increases. For example, when we increase k from 10 to 50, *representativeness* increases from 0.87 to 0.96 for SD-Clu, which means that the absolute difference between $adjCovRate$ of ζ and $covRate$ of R over all instances successively

Table 4.2: **Representative rules (`HepStea`).** Proxy rules for $k = 5$ on the `HepStea` sample and the positive value of the target variable.

#	Antecedent of proxy rule	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 $\wedge$ apolipoprotein A1 $\leq$ 1.56 g/l	0.37	0.41	1.90
4	physical health score $\leq$ 47.3 $\wedge$ increased waist circumference = TRUE	0.29	0.48	2.12
5	high-sensitivity C-reactive protein $>$ 2.8 mg/l $\wedge$ uric acid $>$ 246 µmol/l	0.29	0.46	2.21
SD-Map				
1	diastolic blood pressure $>$ 79.75 mmHg $\wedge$ hip circumference $>$ 98.05 cm $\wedge$ cholesterol-HDL-quotient $>$ 3.015	0.67	0.40	1.85
2	increased waist circumference = TRUE $\wedge$ diastolic blood pressure $>$ 79.75 mmHg $\wedge$ body mass index $>$ 26.3 kg/m^2	0.58	0.45	2.07
3	waist circumference $>$ 88.15 cm $\wedge$ diastolic blood pressure $>$ 79.75 mmHg $\wedge$ body mass index $>$ 26.3 kg/m^2	0.59	0.46	2.11
4	body mass index $>$ 26.3 kg/m^2 $\wedge$ alanin-aminotransferase $>$ 0.385 µkatal/l $\wedge$ diastolic blood pressure $>$ 79.75 mmHg	0.55	0.48	2.20
5	body mass index $>$ 26.3 kg/m^2 $\wedge$ uric acid $>$ 278.5 µmol/l $\wedge$ alanin-aminotransferase $>$ 0.385 µkatal/l $\wedge$ treated urinary tract diseases = TRUE	0.54	0.46	2.34

decreases. Further, for a given k, the representative rules of the baselines are less representative than SD-Clu's proxy rules, e.g., 0.91, 0.92, and 0.91 vs. 0.96 for $k = 50$, respectively (cf. 5th column of plot matrix in Figure 4.4).

4.5 Discussion of Findings

Tables 4.2 and 4.3 show the antecedent, support, and confidence of 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 medical tests (ALAT and LDL). Furthermore, apolipoprotein A1 (ApoA1), a major protein component of high-density lipoprotein (HDL) particles in plasma, is associated with target variable in elderly patients (see fourth hotspot rule in Table 4.2). Lipoprotein metabolism is considered the main process contributing to the development of fatty liver [148]. Besides, Poynard et al. [216] 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. [180] reported increased ultra-sensitive CRP levels in subjects with hepatic steatosis independent of other metabolic states. Similarly, Keenan et al. [152] 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, body mass index and angiotensin II receptor blocker

Table 4.3: **Representative rules (`Goiter`).** Proxy rules for k=5 on the `Goiter` sample and the positive value of the target variable.

#	Antecedent of proxy rule	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

intake (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 [258] as well as subclinical hypothyroidism [89, 270]. The condition of the third HotSpot rule describes the duration of an ECG phase. Jabbar et al. [146] summarized that pathological thyroid hormone levels increase the risk of cardiovascular disease. This association appears to be especially true in the elderly [76]. The fourth rule suggests that certain thrombocyte levels indicate increased thyroid volume, which confirms Erikci et al. [70], who found that hypothyroid patients had higher platelet volume and platelet distribution width than a control group. This shows that our approach delivers relevant results to the application field; here two chronic reversible conditions.

4.6 Conclusion

We have proposed SD-Clu, an algorithm for rule clustering, identifying cluster-representative rules ("proxy rules"), and minimizing the proxy rules shown to the domain expert. Our algorithm 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 subpopulations, 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 fewer rules than the total rule set and (ii) is more representative compared to the baseline approaches, respectively. SD-Clu can not only be applied for subpopulation discovery in epidemiological data but high-dimensional medical datasets in general. In the absence of a

concrete target variable, unsupervised methods for detecting subpopulations (or phenotypes) are needed. Furthermore, more sophisticated visualizations are needed if subpopulations are characterized by many variables rather than just a few. We present our solution to these challenges in the following chapter.

Chapter 5

Visual Identification of Informative Features

Brief Chapter Summary

Knowledge of different disease phenotypes can help understand (a) which patient subpopulations seek treatment and (b) the response to treatment within each subgroup. We present a workflow to (i) determine distinct phenotypes of medical conditions in high-dimensional data, (ii) visualize these phenotypes to explore and compare essential subpopulation characteristics, and (iii) interactively inspect them and their change over time with an interactive web application. We evaluate our workflow on the CHA data by identifying four distinct phenotypes of tinnitus patients.

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 (2020), pp. 1-10. DOI: 10.1038/s41598-020-73402-8.

The supervised methods for subpopulation discovery described in the previous chapters show great potential for applications with one or a small number of well-defined target variables. However, some medical conditions are multifaceted and are not well understood yet. For example, chronic tinnitus is a complex, multifactorial and heterogeneous disorder [45]. Clinical assessment and selection of a suitable treatment is difficult because not all patients benefit equally from each type of intervention [124, 262, 269]. Due to the large number and heterogeneity of available assessment tools, *phenotyping* appears promising to (a) describe the subpopulations of patients seeking treatment and (b) understand the response to treatment within each subpopulation [167, 269, 236]. Clinical acceptance of these empirical results can be further strengthened by comprehensive visualizations that intuitively illustrate each phenotype's major characteristics and differences between multiple phenotypes.

This chapter describes a workflow to

- determine distinct phenotypes of medical conditions with a parameter-free clustering algorithm in high-dimensional data,
- visualize these phenotypes to explore and compare essential subpopulation characteristics, and
- inspect them further in an interactive web application.

Our workflow's effectiveness is demonstrated on the CHA dataset (Section 2.2.2) by exploring phenotypes of tinnitus patients.

This chapter is organized as follows. Section 5.1 describes the clinical value of patient stratification, briefly reviews previous approaches for the example disorder, and discusses significant challenges for phenotyping in high-dimensional data. We present all our workflow components in Section 5.2, namely the clustering algorithm, our novel visualization types, and our web application. Section 5.3 lists the features used for clustering and provides details on the CHA data subset used for workflow evaluation. In Section 5.4, we present the discovered tinnitus phenotypes and describe their characteristics.
In Section 5.5, we discuss our findings from the medical perspective and compare them to related work. Finally, we conclude the chapter with a summary and outlook in Section 5.6.

5.1 Motivation and Comparison to Related Work

Challenges for management and treatment of tinnitus are primarily caused by its clinical heterogeneity, including individual perception, risk factors, comorbidities, degrees of perceived stress, and treatment response (cf. Section 2.2.2). 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 from which all patients equally benefit. The awareness of the existence of distinct patient subgroups may stimulate the development of more effective therapy modules [91]. 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 [167], a combination of features extracted from self-reports, audiometry, and psychoacoustics [269], or neuroimaging data and socio-demographics [236]. Although each of these studies provided insights into 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 in high-dimensional data.

Addressing this requirement, Schlee et al. [237] proposed a compact radar graph to compare of 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 regarding measurements of tinnitus distress and associated comorbidities. However, they did not aim to visualize clustering results. Instead, they 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.

Challenges for clustering in high-dimensional data. There are several challenges when clustering on high-dimensional medical datasets:

- How can we determine an appropriate number of clusters in the absence of a ground truth?
- How can we represent high-dimensional data compactly but also faithfully?
- How can we visualize essential characteristics of high-dimensional clusters?

We present related work for each challenge below.

Determination of an appropriate number of clusters in the absence of a ground truth. 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 k values, cf. Section 6.3.1 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 the 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 (see Section 4.2.1), which favors clusterings that assign similar objects to the same cluster and dissimilar objects to a different cluster. Instead of evaluating clustering quality post-hoc, we decided to leverage an algorithm that automatically determines a suitable number of subgroups already *during* clustering.

Dimensionality reduction. Dimensionality reduction (DR) techniques are often used to project the original high-dimensional data onto a low-dimensional projection that allows simple visualization types such as scatterplots to be used. Ideally, the DR projection preserves the original data's essential structures, such as relative pairwise distance, clusters, outliers, and correlations. Principal Component Analysis [140] (PCA) is a seminal DR algorithm that generates linear, orthogonal combinations of the original dimensions. Each new dimension, called principal component, contains a loading indicating how much variability of the data it covers. Typically, the first two or three dimensions that carry the highest loads are selected for visualization. PCA is not robust to outliers and cannot capture nonlinear relationships. Multidimensional scaling [101] (MDS) is another early DR technique that emphasizes the preservation of pairwise distances, i.e., objects 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 [185] (t-SNE) and Uniform Manifold Approximation and Projection [191] (UMAP) are nonlinear dimensionality reduction methods that represent a matrix of pairwise similarities. The idea is to perserve 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 degree of preservation or contribution of the original features cannot be measured. 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. Medical researchers are usually interested in understanding the relationship between the original variables and the new dimensions to draw actionable conclusions from the patterns found in the projected space. Since the semantics of the original dimensions are lost in DR, we prefer to maintain the high-dimensional space for clustering and visualization.

Visualization of high-dimensional clusters. 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 [142, 155]. 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 merely adding transparency or colors to points to more sophisticated density contours, hexagon binning, layers with aggregated geometric features (minimal spanning trees, alpha shapes, convex hulls), animation, or combinations of several techniques such as splatterplots [190]. However, SPLOMs and other traditional visualization techniques such as parallel coordinate graphs [117] are still more suitable for low-dimensional data. In general, the focus of phenotype visualization is not to represent the specifics of individual subjects but to show the most important general characteristics of each subpopulation. Therefore, to represent patterns in high-dimensional space, we do not create multiple, often cluttered subplots, as the above approaches do, but we represent essential phenotype characteristics in a single visualization. We further provide a web application to explore one and juxtapose multiple phenotypes interactively.

5.2 Discovery and Visualization of Phenotypes

We propose a workflow for determining, visualizing, and inspecting essential phenotypes for a medical condition in high-dimensional datasets. In the following, we present the individual steps of our workflow. Section 5.2.1 describes the clustering algorithm X-means, which internally determines an appropriate number of phenotypes. Then, we present our solution for visualizing these phenotypes with radial bar graphs and radar graphs in Section 5.2.2. Section 5.2.3 describes our web application which combines these visualizations with interactive elements for further phenotype inspection and comparison.

5.2.1 X-means Clustering

X-means [211] is a parameter-free adaption of the popular k-means algorithm, which incorporates the Bayesian information criterion [242] (BIC) to find a good tradeoff between a 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

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

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 the 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 new partitioning's BIC score 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 [144]. Since we did not aim to restrict the solution space with respect to the number of clusters, we set K_{lower} to 2, i.e., the lowest possible value, and we did not set K_{upper}.

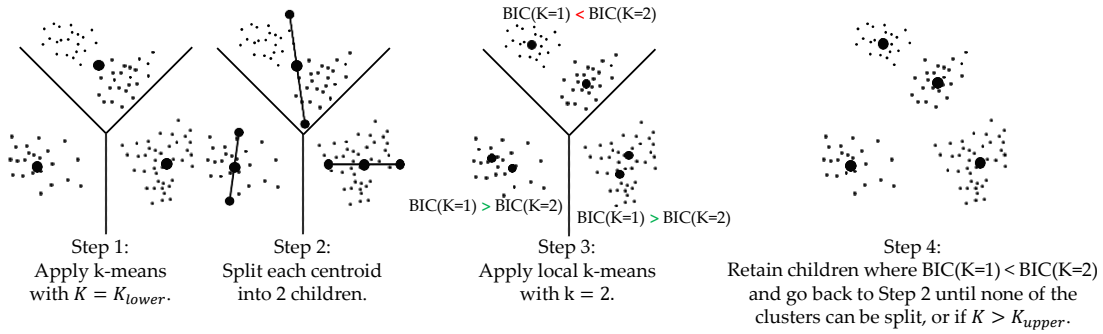


Figure 5.1: **Principal steps of X-means (simplified).** The figure is adapted from [211].

Stability of the clustering result. Like its predecessor, X-means is also a non-deterministic algorithm since the initial centroids' positions are set randomly, leading to different clustering results. To assess the stability the total cluster number, we performed an internal validation where we recorded the number of clusters generated by X-means on 500 bootstrap samples.

5.2.2 Phenotype Visualization with Radial Bar Graphs and Radar Graphs

Requirements for phenotype visualization. Together with domain experts, we transformed the visualization challenges (recall Section 5.1) into the following requirements:

- represent high-dimensional data with dozens of features,
- preserve the semantics of the original features,
- allow for a comparison of multiple clusters at a glance, and
- contrast cluster characteristics with the overall patient mean.

Following these requirements, we implemented (a) a radial bar graph as a visualization of a single cluster (Figure 5.2) and (b) a radial line graph visualization for comparing multiple

clusters at once (Figure 5.3). The radial bar graph 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", representing the feature means of the entire population, i.e., all subjects used for clustering. Because features are standardized (i.e., z-scored) before 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 one 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 can be grouped into (expert-defined) categories, displayed in the inner circle along with the cluster name and the number of subjects assigned.

The radial line graph (Figure 5.3) allows a comparison of all clusters in a single display. Instead of bars, the feature averages are represented as points. Line segments connect points of the same cluster and feature category. Points and line segments are colored according to their respective cluster.

5.2.3 Interactive Exploration of Phenotypes

We provide an interactive application with phenotype visualizations as a web application[1] (Figure 5.4). The following interactive components have been added to the visualizations described in Section 5.2.2. Hovering over a bar, line, or feature label opens a tooltip with additional cluster summaries and compact feature descriptions. Clicking on a bar or label triggers an additional plot showing the original (unscaled) distribution of the respective feature stratified by cluster and, if selected, after treatment. For continuous features, semi-transparent boxplots are placed on top of violin plot [131] layers. For categorical features, points and error lines show the percentages of each category and the 95% confidence interval. While clustering is performed on static data, we added an option to display also indicators of temporal change, which helps, for example, to discover potential differences in response to treatment between the phenotypes visually. To this end, we extended the radial bar graph by showing cluster averages at two time points (T0 and T1) with adjacent bars. A line connects the ends of a pair of bars with an arrowhead pointing from the T0 score to the T1 score. The connecting lines and feature labels are colored according to their relative value change from T0 to T1. Given the user-defined parameter Δ, i.e., the minimum relative difference between T0 and T1 considered as a change, elements are colored

- red if the T1 score is greater than the T0 score by at least Δ,
- green if the T1 score is smaller than the T0 score by at least Δ, and
- black otherwise.

5.3 Selection of Measurement Instruments

Discussions with tinnitus experts about the selection of measurement instruments (hereafter denoted as *features*) for clustering the CHA data resulted in two main requirements: (1) The

[1] A demo is available at https://unmnn.de/phs/app/.

selected features should cover the clinical heterogeneity of tinnitus to a high degree. (2) If available, more robust compound scores should be preferred over single items from a questionnaire.

From the routine questionnaire assessment battery (cf. Section 2.2.2), we selected a total of 64 features[2] 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, pain experiences, somatic expressions, affective symptoms, tinnitus-related distress, internal resources, perceived stress, and mental quality of life.

From a total of 4103 patients, data from 2875 (70.1%) was incomplete and therefore excluded. The N = 1228 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 (80.5%), post-treatment data were also available and used to explore treatment effect differences between clusters visually. Since most of the features have greater scores for higher health burden, we reversed the remaining features with greater scores for a 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 scaling. A feature X with expected value $E(X) = \mu$ and variance $Var(X) = \sigma^2$ was standardized into $Z = \frac{X-\mu}{\sigma}$. For Z, it holds that $\mu = 0$ and $\sigma^2 = 1$.

5.4 Results

According to X-means, four clusters (also referred to as phenotypes hereafter) represent an optimal solution for the CHA data. This result is confirmed by bootstrap validation: Figure 5.5 shows that four clusters are formed most frequently (82 times; 16.4%) among the 500 runs.

Phenotype 1 (PT 1) represents the largest subgroup (697 of 1228 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 (b), (e), (i), and (g)).

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 a high tinnitus burden besides 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 severely 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 percentage 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 complete list of features is provided in Appendix A.

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 features 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 percentage 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 four subgroups (mean 47.3 years), with the largest share of male patients (Figure 5.6 (c)).

Figure 5.7 depicts the top 10 features with the greatest 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 feature is ADSL_depression with an average difference between T1 and T0 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.

5.5 Discussion of Findings

5.5.1 Juxtaposition of the Phenotypes

Our clusters comparison showed that some questionnaires and characteristics differed considerably between patient phenotypes. In particular, patient subgroups differed substantially in coping behaviors, stress, tinnitus burden, perceived pain, discomfort, and perception of life quality. In contrast, patients did not appear to differ concerning localization and noise. Regarding the separability between phenotypes, the predominantly high correlations between features 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, among others.

5.5.2 Interpretation of the Phenotypes from the Medical Perspective

We discussed the phenotypes' clinical relevance with three of the five tinnitus experts who co-authored the original publication [201].

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 these patients' biased presentation (*"everything is fine if it were not 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*) represents a patient subgroup characterized by somatopsychic symptom expression, i.e., physical symptoms that may reflect stress or underlying medical conditions. To meet this patient subgroup's needs, multimodal interventions may include a proportion of body-oriented procedures such as relaxation exercises or physiotherapy. However, 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 (*distress group*) reported above-average perceived stress, accompanied by physical exhaustion and anxious-depressed mood. This group includes younger, more employed, and more male patients who reported chronic distress and may be susceptible to burnout syndrome with subjectively reduced mental performance ("hamster wheel"), which describes 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. Like 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).

5.5.3 Comparison to Related Work on Tinnitus Phenotypes

Without a "ground truth" and given that different sets of available measurements were used, it is difficult to compare our results with similar studies. An advantage of our approach is the inclusion of a wide range of self-report questionnaire assessments. Other studies also used audiometry [269, 167], and cardiac imaging data [236]. PT 2 (psychosomatic distress group) seems to be associated with the "constant distressing tinnitus" subgroup reported by Tyler et al. [269], as the mean scores of tinnitus-related health distress were much larger than in the other subgroups. Obviously, selecting a meaningful set of characteristics is central to the effectiveness of any cluster analysis.

The closest to our radial bar graph visualization is the radar graph proposed by Schlee et al. [237], 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 [237] depend heavily on the arrangement of features since the primary criterion for comparison is the polygons' shapes. 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. 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, making it easier to find features and compare similar features.

5.6 Conclusion

We have presented a workflow for determining, visualizing, and inspecting essential phenotypes of a medical condition in high-dimensional datasets using the example of tinnitus. Although we have demonstrated our workflow's usefulness on a specific disorder, it can be easily adapted to any other medical condition.

To reduce the amount of input necessary from the medical expert, we leverage a parameter-free clustering algorithm for phenotype discovery. Although the optimal number of clusters is four for our dataset, 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 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. Instead of using X-means, we could have evaluated k-means with different k using a cluster quality function. The suitability of such a more heuristic approach is investigated in Chapter 7.

Our novel visualization types provide the medical expert with a quick overview of the most important characteristics of and differences between subpopulations. These are integrated into and a web application with interactive functionalities for cluster inspection and juxtaposition. Both the visualization and the application are not tinnitus-specific but can be used to display a compact summary for any condition or subset of index symptoms. Whether clinicians will adopt the visualizations to guide appropriate tinnitus management strategies 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.

In Chapter 8, we revisit the CHA data and focus on supervised subpopulation discovery and post-hoc interpretation of classification models by the example of tinnitus-related distress and depression, respectively. In future work, we will validate these phenotypes on a different cohort of tinnitus patients. Furthermore, we will explore other clustering algorithm families. For example, *self-organizing maps* [158] seem to be a suitable candidate, as they already offer dedicated visualizations [289] out of the box, and their extension for timestamped data [234] may allow us to study phenotype changes over time.

By excluding patients who did not complete all questionnaires at T0, there may be a selection bias. Possible reasons for non-completion 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 these questionnaires' contributions 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, the next step would be to investigate the effects of treatment on these phenotypes in more detail and determine whether some patient phenotypes benefit more than others.

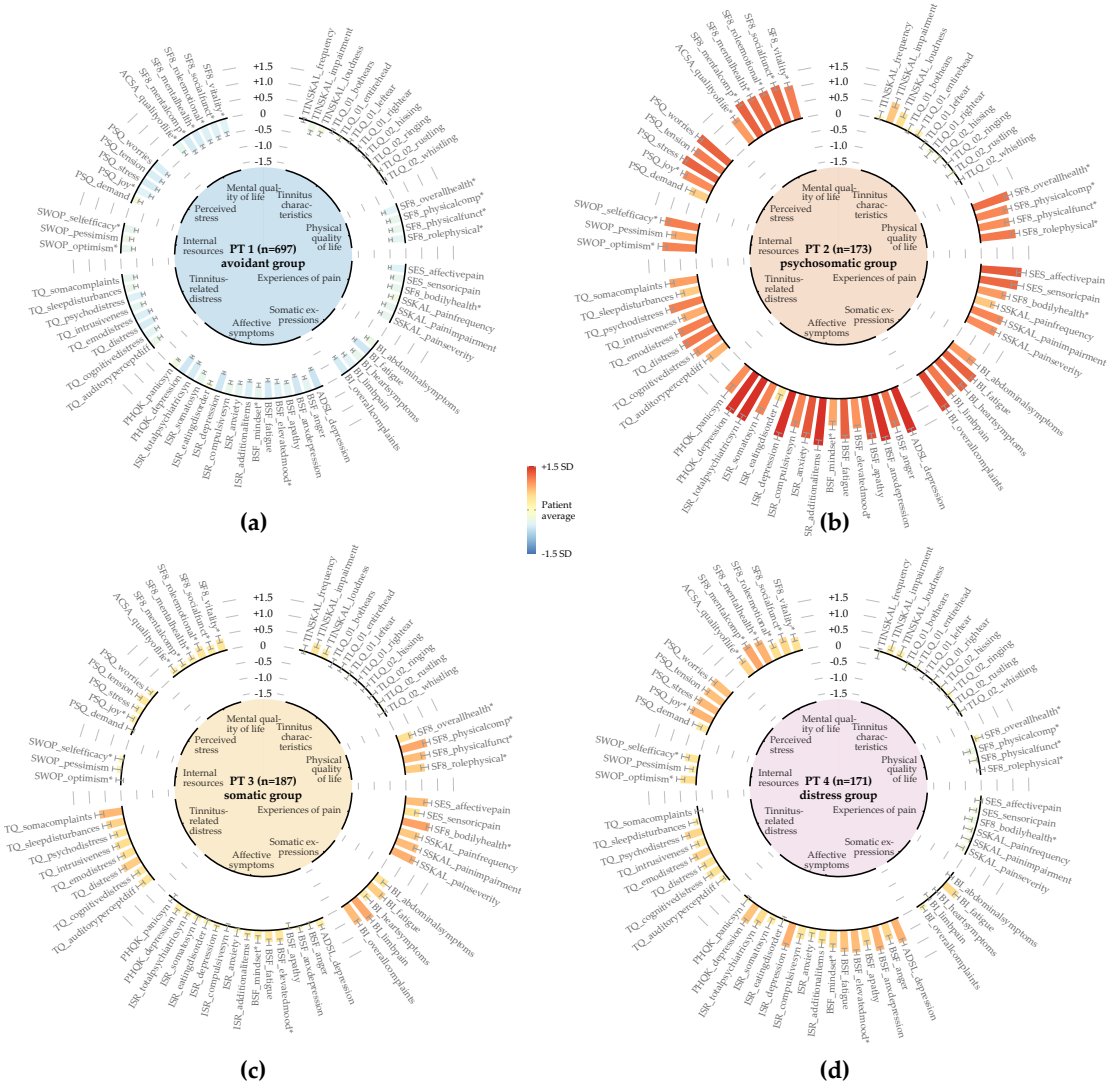


Figure 5.2: **Radial bar graphs for each of the four phenotypes (PT).** (a) PT 1 characterizes the subpopulation with the lowest health burden. (b) PT 2 includes the most suffering subjects, in whom all mean values of psychosomatic and somatic characteristics exceed the population mean by more than 0.5 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 bar's height and direction 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 nine categories defined by tinnitus experts. The categories are shown inside the inner circle. See Appendix A for a feature description. The figure is adapted from [201].

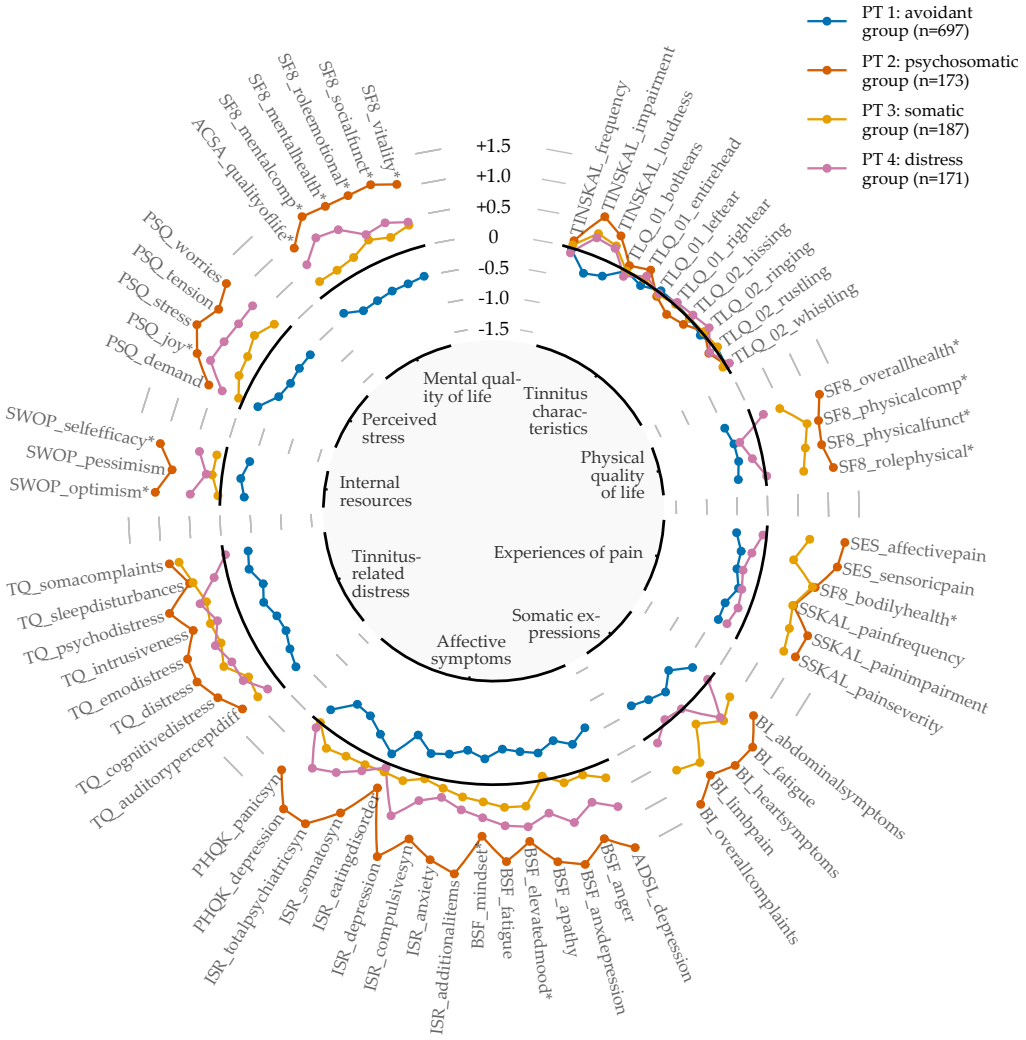


Figure 5.3: **Radial line graph for phenotypes comparison.** Points show within-phenotype feature averages. Points depicting features of the same category are connected with line segments. Points and lines are colored by cluster. See Figure 5.2 for a description of the phenotypes and Appendix A for a description of the features. The figure is adapted from [201].

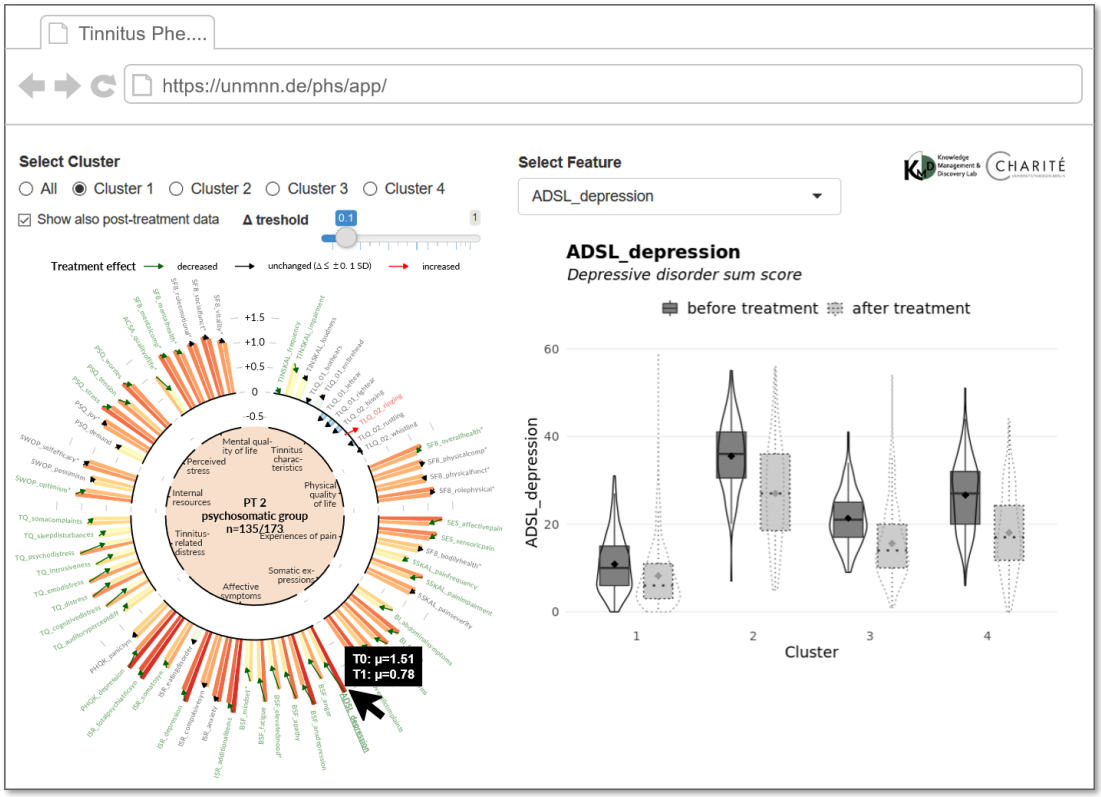


Figure 5.4: **The user interface of the phenotype exploration application.** Interactive components enhance the radial bar graphs: hovering over a bar or feature label opens a tooltip with additional cluster summaries and a compact feature description. Clicking on a feature updates the right plot showing 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 [131] layers. In contrast, for nominal features, category proportions alongside their 95% confidence intervals are displayed as points and error lines, respectively.

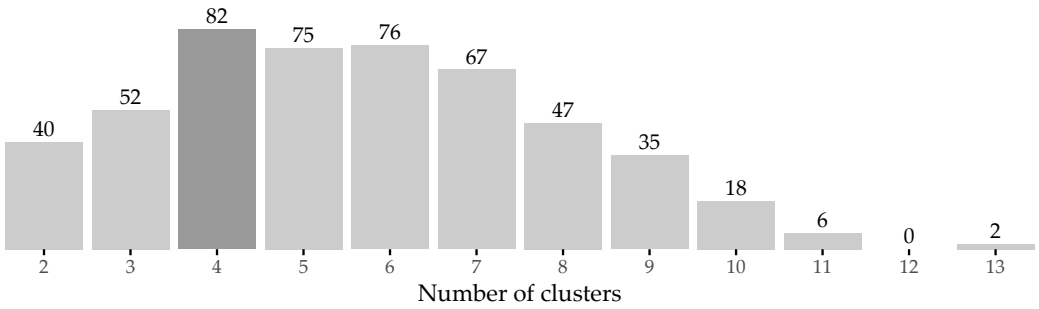


Figure 5.5: **Results of internal validation.** Bars show the frequency of the number of clusters generated by X-means for 500 bootstrap samples. The most common cluster number was 4, which occurred in 82 samples (16.4%).

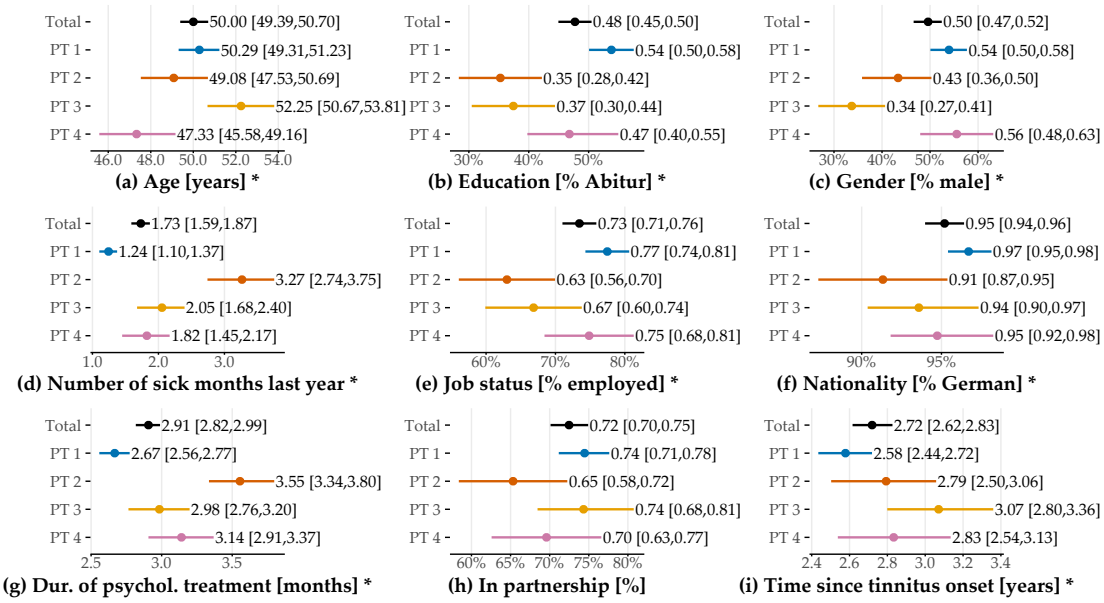


Figure 5.6: **Inter-phenotype comparison of demographic characteristics.** Summaries are given as *mean [95% confidence interval]* for the entire population and each of the four phenotype subpopulations. Confidence intervals were estimated using nonparametric Basic Bootstrap Sampling [58] with 2000 samples each. The Kruskal-Wallis test was used to compare differences between phenotypes for continuous features (such as age), and Pearson's chi-square test was used for categorical features (such as gender). An asterisk indicates statistical significance ($\alpha = 0.05$). Correction for multiple comparisons was not performed due our approach's exploratory nature. The figure is adapted from [201].

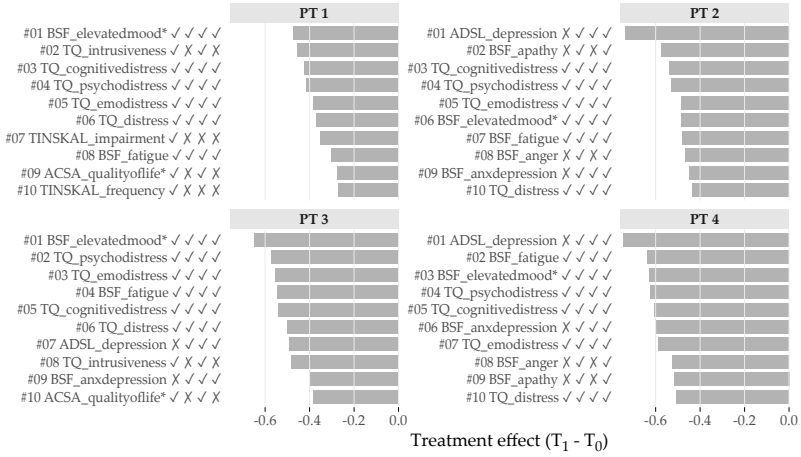


Figure 5.7: **Cluster-specific top 10 features with highest average treatment effect magnitude.** Bars depict the intra-cluster average differences between the measurements at T1 and T0 based on the standardized values. Lower values represent better treatment effectiveness. The symbols right to the feature 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.

Part II

EXPLOITING DYNAMICS

Chapter 6

Constructing Evolution Features to Capture Change over Time

Brief Chapter Summary

We propose a framework to extract "evolution features" from timestamped medical data, which describe the study participants' change over time. We show that exploiting these novel features improves classification performance by validating our workflow on the SHIP data for the target variable hepatic steatosis.

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: *Computer-Based Medical Systems (CBMS)*. 2015, pp. 121-126. DOI: 10.1109/CBMS.2015.12.

Medical studies with a longitudinal design collect participant data from questionnaires, medical examinations, laboratory analyses, and imaging repeatedly over time [282, 143, 136]. Hidden temporal information could be made explicit by constructing features that describe the subjects' change over time. However, there is a lack of applicable methods for timestamped data with a tiny (< 5) number of moments. We present a framework addressing this shortcoming and demonstrate that augmenting the feature space with so-called *evolution features* increases classification performance and yields understandable descriptors of change worthy of further investigation.

Section 6.1 describes work related to the construction of temporal representations in medical data. Section 6.2 introduces the notation and problem formulation. Section 6.3 presents our evolution feature framework, including a full workflow that encompasses steps to extract evolution features, dealing with class imbalance, and feature selection. Section 6.4 describe the evaluation setup. We report on our results in Section 6.5 and present evolution features found to be important for class separation in Section 6.6. We conclude this chapter in Section 6.7.

6.1 Motivation and Comparison to Related Work

Epidemiological studies serve as a basis for the identification of risk factors associated with a medical condition [106, 30, 206]. Machine learning is still relatively 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 [14] and the discovery of factors (including biomarkers) that modulate a medical outcome [223, 272]. 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.

From a technical point of view, classification problems on timestamped data can be very different. They can be broadly summarized into three categories:

a. Each instance (patient, study participant) has a label at each time point t. The goal is to assign the instance's label at $t + 1$. Typically, this is a *data stream classification* problem [271, 3].
b. Each instance is associated with only one label for the all time points. Typically, this is a *timeseries classification* problem [74, 141].
c. There are no labels but the time point of an *event* is to be predicted. This problem is concerned with *event prediction* [303, 16, 27].

Our classification problem falls under (a), but we cannot use the advances of stream classification because we our data is unlabeled except for the last time point and our stream of three time points is tiny.

In clinical applications, temporal information is often exploited, predominantly for the analysis of patient records. For example, Pechenizkiy et al. [210] analyzed streams of recordings to predict rehospitalization because of heart failure events during remote patient management. Sun et al. [256] computed the similarity between streams of patients from patient monitoring data. Combi et al. [55] reported on streams of life signals, particularly 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 a skewed class distribution. In a *longitudinal* study of this kind, recordings for the same cohort member are made at each moment. Hielscher et al. [127] 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. Afterward, original and sequence features are used in conjunction for classification. Hielscher et al. [127] 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. [248]. They proposed a method that predicts patient evolution from timestamped data by clustering them on similarity and predicting cluster movement in the multi-dimensional space. However, the patient data considered in [248] 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) [282], recall Section 2.2.1. For the

SHIP cohort, the assessments (interviews, medical tests, etc.) were recorded in several *moments* (SHIP-0, SHIP-1, etc.), 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 months or even years. Hence, measurements of the same assessment in an epidemiological dataset are few and possibly far apart. The large period between two consecutive recordings complicates applying 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 five years ago irrelevant for learning the participant's current health state. Another patient may have no lifestyle changes and no illnesses, so their past data reflect only aging.

A further challenge is that the label is unavailable for several waves. 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. Consequently, for a given SHIP participant, the 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 – using a growth model [249] – 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 with similar recordings at one of the three moments. Then, we connect the clusters across time, thus capturing each cluster's transition from one study wave to the next. These transitions reflect the evolution of subpopulations, 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. We show that this new, augmented dataset, combining labeled and unlabeled data on individuals and subpopulations, improves classification and delivers additional insights on some factors associated with hepatic steatosis.

6.2 Problem Formulation

We now provide some basic symbols and essential functions for our classification problem, summarized in Table 6.1.

Our longitudinal study comprises data for a participant $x \in X$ for every feature $f \in F$ and every time point $t \in \{1, \dots, T\}$ (also referred to as moment). The (single) value of x for feature f at moment t is denoted as $v(x, f, t)$; the set of all values for x at t, i.e., $\{v(x, f, t) | \forall f \in F\}$, is denoted as $obs(x, F, t)$. For some participant, $v(\cdot)$ is NULL, i.e., there may be missing values. Our goal is to to predict the class label of x at moment T, i.e., $label(x, T)$ by using data from all observations in all previous moments $t \in \{1 \dots T-1\}$. The class label of each participant is unavailable except for $t = T$. We model this prediction task as conventional classification problem. We aim to improve classification performance by expanding the feature space with change descriptors, the so-called *evolution features*, presented next.

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

Term	Description
x	a study participant from cohort X
t	a study moment, one of $\{1, \ldots, T\}$
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., $\{v(x, f, t) \vert \forall f \in F\}$
$label(x, t)$	the class label of x at moment t
$Z(t)$	all observations at t, i.e., $\{obs(x, F, t) \vert \forall x \in X\}$
$c(x, t)$	cluster membership of x at t; if x is an outlier at t, then $c(x, t)$ is NULL
$d(x, z, t)$	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 $c(x, t)$

6.3 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 a 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 clusters' change 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).

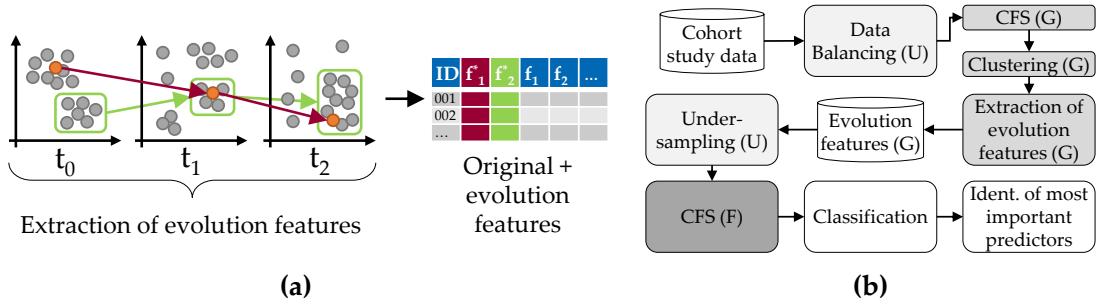


Figure 6.1: **Concept of evolution feature extraction for classification performance improvement**. (a) Clustering of longitudinal cohort data and subsequent generation of evolution features from the change of individuals (red) and whole clusters (green). (b) Overview of the classification workflow.

In the following, we describe the clustering of study participants (Section 6.3.1), the generation of evolution features (Section 6.3.2), and our feature selection strategy to extract a subset of *informative* features as input for classification (Section 6.3.3), while dealing with class imbalance by undersampling the majority class.

6.3.1 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 [72] 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).

Distance function. For the distance between participants x, z at t, we use the *adjusted heterogeneous Euclidean overlap metric* [127, 294], 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 Ester et al. [72], which determines a suitable *eps* value for a given *minPts* value, 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 kth 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 the maximum distance between g and l.

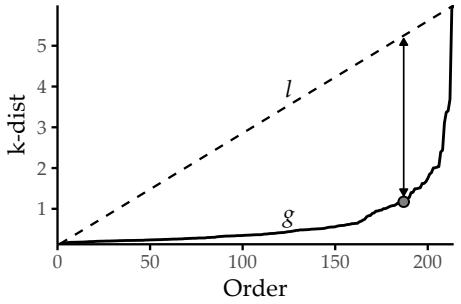


Figure 6.2: **Setting *eps* based on the k-dist graph for a given *minPts*.** The k-dist graph g depicts the sorted distances to the points' kth next neighbors. A suitable eps_{opt} can be identified with the maximum distance between the k-dist and the line l that connects the first and the last point 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.

6.3.2 Constructing Evolution Features

We extract information about the participants' evolution, 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. Table 6.2

describes all evolution features. They are divided into four categories: (A) features for each participant at each moment, (B) evolution features describing some aspect of change between study waves for an individual participant, (C) evolution features measuring changes between study waves concerning feature values for each participant, and (D) evolution features linked to a whole cluster.

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 [259] and (5) Silhouette coefficient [259] of x, and (6) the (graph-based) separation [259] 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).

6.3.3 Extending Correlation-Based Feature Selection to Include Evolution Features

For imbalanced data, feature selection and classification are often biased in favor of the majority class [177]. Hence, before feature selection, we undersample the majority class and generate a balanced data set to select features informative with respect to all classes. We use the feature selection method of Hielscher et al. [128] as follows: we invoke correlation-based feature selection [112] (CFS), which builds up a feature set by iteratively inserting the feature that adds the 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 Fayyad and Irani [75]. 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, which 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.

6.4 Evaluation Setup

We evaluate our workflow with 10-fold cross-validation on four off-the-shelf classification algorithms: random forest [35] (RF), C4.5 decision tree [221], Naïve Bayes (NB), and k-nearest neighbor (kNN). First, we compare each algorithm's generalization performance when used alone (baseline variant) vs. when incorporated into our workflow (workflow-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`).

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

#	Name	Description
(A) 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))`	a is one of cohesion, Silhouette, separation [259]; 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
(B) 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'`	Four-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 $d(x, \widehat{c(x,t)}, t') - (x, \widehat{c(x,t)}, t)$
(C) Evolution features associated with 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 $v(x,f,t)$
(D) Evolution features linked to a whole cluster c		
20	`smaller_Cluster_Fraction_t_t'`	Difference of the size of cluster c at t' and its size at t; c is matched to the clusters of t based on 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`	✗	✗	✗

Sensitivity (true positive rate), specificity (true negative rate), and F-measure (harmonic mean of precision and recall) serve as evaluation measures. As shown in Table 6.3, `Baseline` invokes only the classification algorithm; we use the classification algorithm 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 list `U-G` and `U--` explicitly.

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 6.3.1) 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 3 and 4 on the differences between female and male participants with respect to the outcome, we run the experiments on the whole dataset (`PartitionAll`) and on the partitions of female (`PartitionF`) and male (`PartitionM`) participants. Finally, we list the most important features found in `PartitionAll` and its two subsets.

6.5 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 in specificity. The workflow-enhanced Naive Bayes performs best concerning sensitivity for any k and best for k = 31. Decision trees exhibit the highest F-measure, with improvements on sensitivity and F-measure compared to its simple variant, albeit specificity being slightly 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 the right part of Figure 6.3). One explanation for the relatively low sensitivity of the simple RF variant is the large number of trees (100) learned on data samples containing very few positive examples: RF could have been trapped by the many majority class examples. This result 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 for 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

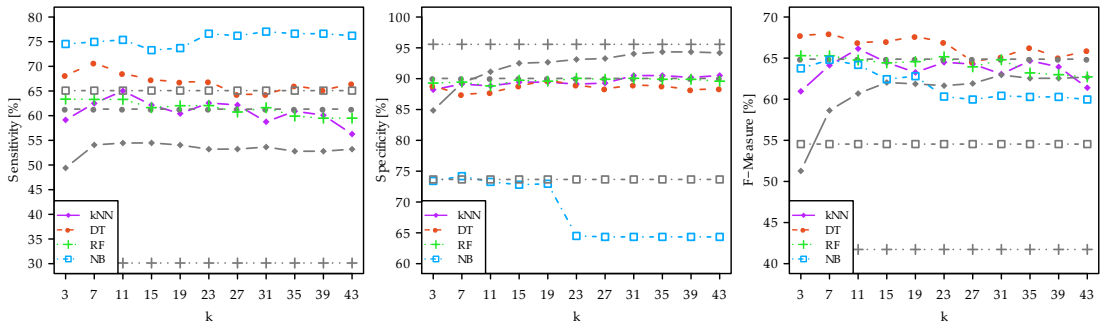


Figure 6.3: **Comparison of classification performance between workflow and baseline.** Sensitivity (left), specificity (center) and F-measure scores (right) of different classifiers when varying the number k of neighbors to a cohort member which impacts the clustering result. For each classifier, two 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. The figure is adapted from [199].

counterparts in terms of sensitivity and F-measure. For some algorithms, our workflow prevents overfitting to the negative class.

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

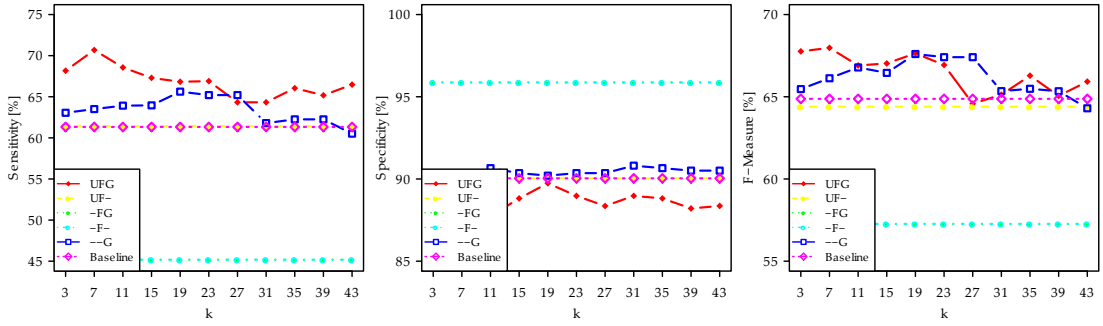


Figure 6.4: **Comparison of classification performance for workflow components.** Sensitivity (left), specificity (center) and F-measure scores (right) for each workflow variant and the baseline using decision tree for learning. The figure is adapted from [199].

6.6 Identification of Important Evolution Features

The performance of the workflow variants, which include feature selection, indicates that a small number of features is sufficient for class separation. Hereafter, we report on the evolution features selected for classification and appearing among the top 15 features according to information gain for each partition; we term these features “important”.

Figure 6.5 shows that for `PartitionAll` 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 belonging to the positive class exhibit a higher median in `separationDelta_g_1_2` than participants without the disorder, 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. On average, study participants from both classes lose weight when they grow older. Negative participants reduce more weight than positive participants (cf. Figure 6.5 (c)), which generally reflects differences in lifestyles. 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 fatty liver participant is also more likely to exhibit the disorder than it is for a participant without fatty liver.

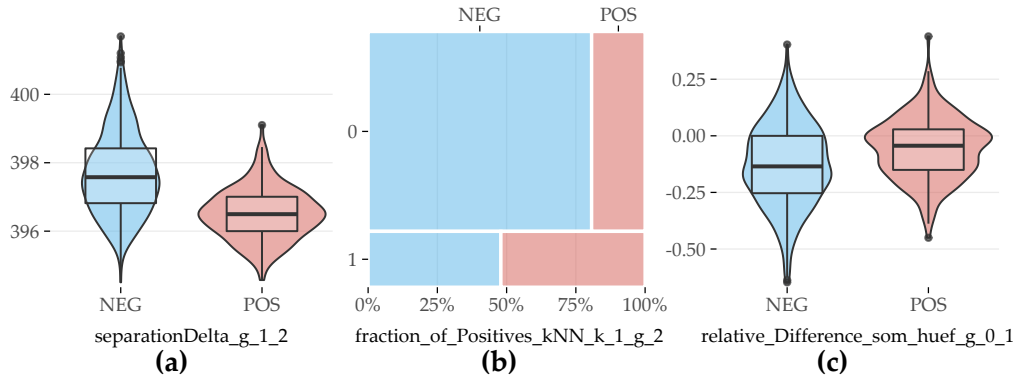


Figure 6.5: **Selected evolution features with the highest contribution to class separation for `PartitionAll`.** (a) Difference in cluster separation between SHIP-1 and SHIP-2. (b) Whether (=1) or not (=0) the nearest neighbor is of the positive class in SHIP-2. (c) Relative difference in hip circumference between SHIP-0 and SHIP-1.

For `PartitionF`, 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).

For `PartitionM`, 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 cluster separation between SHIP-0 and SHIP-1 (#6). For both features, participants exhibiting the disorder have greater values.

6.7 Conclusion

We have extended our set of methods for static medical data presented in Part I by proposing a workflow for the classification of longitudinal cohort study data that exploits inherent temporal information by clustering the cohort participants at each moment, linking the clusters, and tracing participant evolution over the study's moments. We extract *evolution features* from the clusters and their transitions, which are added to the feature space and subsequently used for

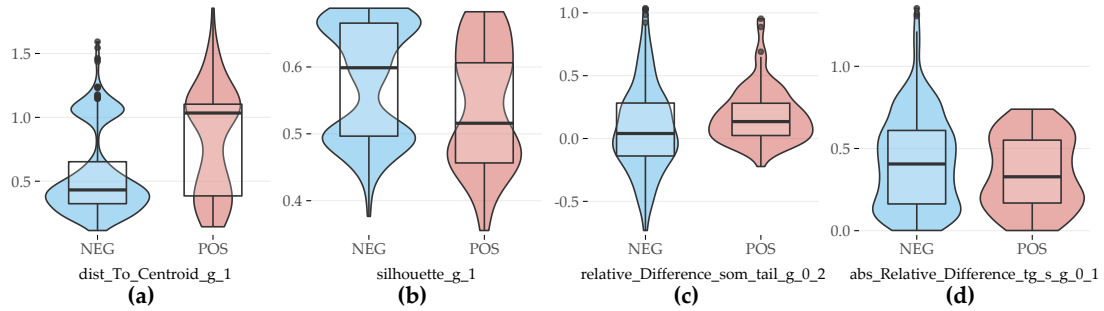


Figure 6.6: **Selected evolution features with the highest contribution to class separation for `PartitionF`.** (a) Distance to cluster centroid in SHIP-1. (b) Silhouette coefficient in SHIP-1. (c) Relative difference in waist circumference between SHIP-0 and SHIP-2. (d) Absolute value of relative difference in serum triglycerides between SHIP-0 and SHIP-1.

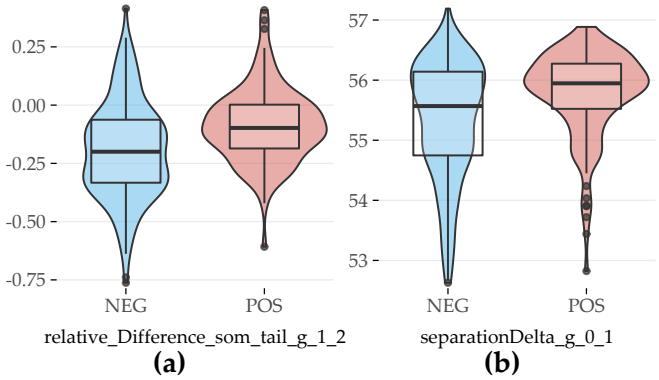


Figure 6.7: **Selected evolution features with the highest contribution to class separation for `PartitionM`.** (a) Relative difference in waist circumference between SHIP-1 and SHIP-2. (b) Difference in cluster separation between SHIP-0 and SHIP-1.

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 used alone and without undersampling the skewed data. We have shown that the change of somatographic variables' values and cluster quality indices over time are predictive.

A limitation concerns the assessment of a feature's importance using information gain. The additional merit a feature has towards model predictions cannot be assessed being decoupled from the actual model. Furthermore, models may incorporate complex feature interactions that are not captured or subpopulation-specific differences in feature importance. Part III addresses *post-hoc model interpretation*, which includes more sophisticated approaches of measuring a feature's attribution to a model.

Chapter 7

Feature Extraction from Short Temporal Sequences for Clustering

Brief Chapter Summary

Embedded mHealth devices continuously record vital functions. These temporal sequences arrive mostly as raw sensor measurements, requiring the extraction of meaningful features. We propose similarity measures for short temporal sequences to build representations of distinct subpopulations. We validate our approach by identifying characteristic plantar pressure profiles from recordings of a controlled experiment with diabetic foot syndrome patients and non-diabetic volunteers.

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: *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* 54.102712 (2020), pp. 1-11. DOI: 10.1016/j.ebiom.2020.102712.

Nowadays, mHealth devices, such as smart wearables, are ubiquitous in everyday life [92]. These embedded systems continuously record and process various vital functions and send feedback to

the user, for example, via popup notifications or dashboard reports. Thus, they are designed to help live healthier by contributing to more effective exercise or a better diet. mHealth solutions are also increasingly being used in clinical settings. Pattern recognition in timestamped mHealth data is used to timely detect adverse health events.

For example, in diabetes patients with sensory neuropathy, repeated excessive pressure loads can aggravate plantar tissue destruction and ultimately lead to foot ulcers, in the worst case, even to amputation [33]. Since frequent clinical examinations would be too costly and bothersome for the patient [195], sensor-equipped systems for the patient's shoe insoles are developed. These systems measure foot pressures and temperatures continuously and transmit recordings wirelessly to a smart device, which processes the data to send offloading instructions when potentially harmful pressures are detected [1] or temperature differences between left and right foot exceed pre-specified thresholds [195]. However, a universal pressure threshold with high sensitivity and specificity with respect to ulcer development has not yet been determined [288, 77]. Besides, when a significant difference in temperature between the left and right foot is noted, it is often already too late to warn the patient, as ulceration has already begun. Evidence suggests that there are subpopulations with distinct plantar pressure patterns requiring different preventive measures, for example, different *peak pressure* thresholds [93, 59, 23, 60].

In the previous chapters, we presented methods for analyzing static data (Chapters 3–5) and timestamped data (Chapter 6) with a small number of time points. Raw mHealth data comprises hundreds or more time points, thus requiring the extraction of meaningful features before any actual data analysis. This chapter proposes similarity measures for short temporal sequences to create representations of distinct subpopulations from mHealth data. The DIAB data described in Section 2.2.3 serve as proof-of-concept validation.

This chapter is organized as follows. Section 7.1 describes the medical background of diabetic foot syndrome and reviews previous subtyping approaches to detect plantar pressure patterns. Section 7.2 presents our approaches for modeling regional plantar pressure similarity and identifying distinct foot profiles by clustering. We report on our results in Section 7.3 and compare them to similar studies in Section 7.4. We conclude the chapter in Section 7.5.

7.1 Medical Background

Diabetic foot syndrome (DFS) has a substantial negative impact on life quality [32]. Affected patients have higher mortality [32, 197] and are at higher risk of foot ulcerations [250]. More than 85% of foot amputations relate to foot ulcers [102, 280].

Peripheral sensory neuropathy is most predisposing for foot problems in diabetes patients [33]. In sensory neuropathy, damaging events and injuries go unnoticed, and continued insults can exacerbate tissue destruction. However, understanding of the underlying pathomechanisms of tissue destruction in the absence of trauma is limited.

Lower extremity (micro)blood flow of diabetes patients was compared with healthy controls to establish a causal relationship between decreased blood flow and the occurrence of diabetic foot syndrome [257, 73, 235, 225, 224, 9]. Persistent elevated plantar pressure is a major risk factor for ulceration in diabetes patients [276, 8, 173]. Custom-made footwear and orthopedic shoes are used for ulcer prevention and individual therapy [86, 43, 228]. However, a pressure threshold with high sensitivity and specificity with respect to ulcer development has not yet been determined [288, 77]. Researchers also investigate how pressure is applied to each foot region. Bennetts et al. [23] suggested differences between foot types and biomechanics that result in differences in pressure distribution between regions of the same foot. Deschamps et al. [60] proposed a stratification of patients based on their plantar pressure pattern homogeneity.

This approach may avoid the pitfall of smoothing away variations within a pathophysiological subgroup.

In the absence of a ground truth, clustering is often used to derive subgroups of patients who share similarities in pressure distribution, usually focusing on *peak plantar pressure*, i.e., the maximum observed pressure recorded for a single measurement, for example, the maximum pressure at a sensor during a step [93, 59, 23, 60]. Giacomozzi and Martelli [93] examined the plantar pressure curves of DFS patients and control subjects. They performed clustering based on the curves' similarity in shape and amplitude. They juxtaposed the clusters to predefined subgroups: For example, all subjects in the group with increased peak pressure and muscle weakness or limited joint mobility were assigned to one cluster [93]. De Cock et al. [59] examined peak pressure measured at different foot regions during jogging and identified four pressure patterns with different pressure centers. For example, the "M2 pattern" is the cluster where the maximum of the mean total regional impulse is located at the second metatarsal bone. Bennetts et al. [23] considered seven plantar regions and formed subgroups of patients with similar peak pressure distributions in these regions. Deschamps et al. [60] focused only on plantar forefoot peak pressure distribution and studied both patients and controls. They formed clusters of patients with similar forefoot pressure distributions and identified a cluster consisting only of diabetes patients.

Similar to [23, 59], we focus on patient stratification with clustering. As in [23, 93], we consider all foot regions, not just the forefoot. In contrast to all the aforementioned methods that determined similarity in a single way and then used k-means for clustering, we present alternative ways of modeling similarity in terms of plantar pressure and we contrast the clusters formed by different clustering algorithms in terms of their quality. Besides, we provide visualization aids for inspecting the most representative instance (patient foot) in each cluster.

Initially, the study experiment was performed exclusively on patients. Volunteer data became available only gradually. Hence, this chapter is divided into two parts: In part A, we present alternative distance measures for regional plantar pressure load and evaluate them by clustering experiment data from diabetes patients. In part B, we use the best distance measure from part A to cluster the subgroups of diabetes patients, healthy volunteers, and a combined group separately.

7.2 Modeling Similarity of Regional Plantar Pressure

Our approach includes feature extraction (Section 7.2.1), plantar pressure distribution similarity modeling (Section 7.2.2), clustering of study participant feet (Section 7.2.3), clustering evaluation (Section 7.2.4), and visualization. We present these steps below.

7.2.1 Feature Extraction from Short Time Series

The pressure measurements for one foot of an example participant are shown in Figure 7.1. For each experiment, each foot, and each of the eight sensors, the pressure recordings result in a time series. Within each time series, *stance episodes*, i.e., phases in which the patient was standing, are indicated by dashed boxes. During these episodes, the example patient applied more pressure to the central sensors MTB-3 and Calcaneus than Digitus-1 and Lateral, suggesting a balanced pressure pattern.

For each session i, i.e., for each experiment, participant and foot separately, the minimum and maximum observed pressure values over all sensors were identified, $r_i^{\min}$ and $r_i^{\max}$, respectively.

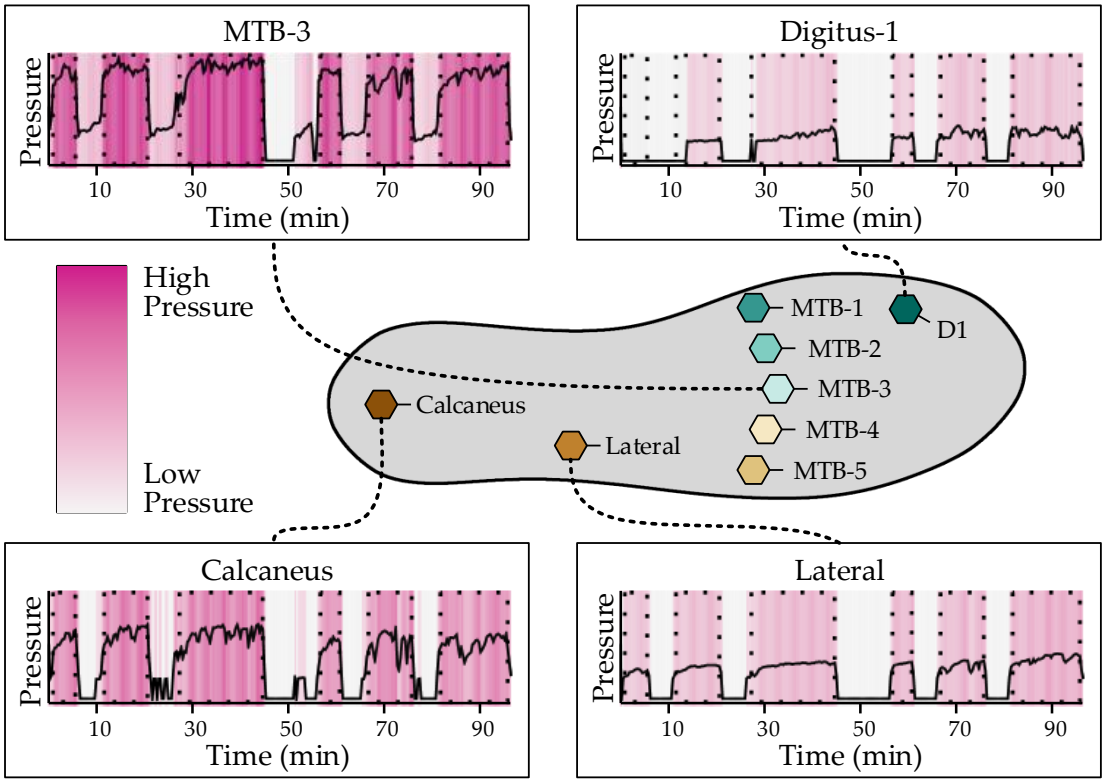


Figure 7.1: **Insole sensor locations and pressure time curve examples.** Sensor locations on the insole (center) and four pressure time curves of representative foot regions of an example patient. Dotted boxes highlight time intervals where the patient was asked to stand and apply pressure. These lasted over 5, 10, and 20 minutes, respectively. D1: Digitus-1; MTB: metatarsal bone. The figure is adapted from [200].

Then, within each session i and for each sensor s each observed pressure value $r_{i,s}$ was normalized into the *relative plantar pressure* (RPP) value

$$r^*_{i,s} = (r_{i,s} - r_i^{\min})/(r_i^{\max} - r_i^{\min}). \tag{7.1}$$

7.2.2 Distance Measures

We consider three alternative ways of modeling plantar pressure similarity:

- similarity based on relative plantar pressure,
- similarity based on pressure distribution in pairs of sensors, and
- similarity based on centers of pressure.

Similarity based on relative plantar pressure. We define the distance d_{RPP} between sessions i, j as Euclidean distance between the mean RPP over the eight plantar regions:

$$d_{\text{RPP}}(i,j) = \sqrt{\sum_s \left(\mu(r^*_{i,s}) - \mu(r^*_{j,s})\right)^2}, \tag{7.2}$$

where $\mu(r^*_{i,s})$ is the average RPP in session i at region s.

Similarity based on pressure distribution in pairs of sensors. For each pair of regions s, t of a session i, a simple linear regression model $X_t = \beta_0 + \beta_1 X_s$ is derived, capturing the linear relationship between the plantar pressure recordings at s and those at t. Figure 7.2 shows an example of the MTB-3 and MTB-4 sensors with three feet. While a generally applied less pressure than b, the linear relationship between the recordings at MTB-3 and MTB-4 is similar, as indicated by the nearly parallel regression fit. Compared to a and b, the regression fit of c has a smaller slope, indicating an increase in pressure at MTB-3 is associated with a smaller increase in pressure at MTB-4. The regression lines also differ in the goodness of fit. For example, the residual values are, on average, smaller for b than for c.

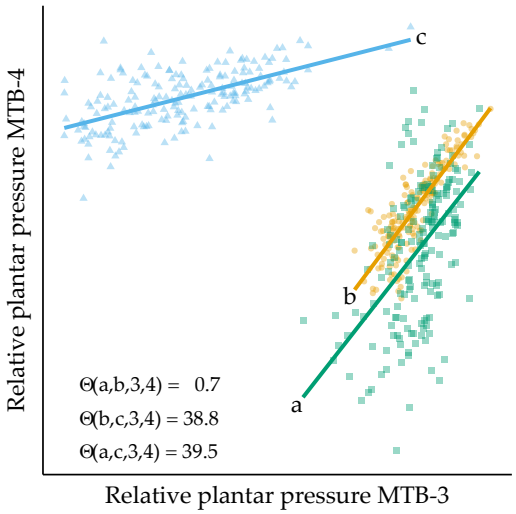


Figure 7.2: **Example of three RPP distributions.** For three example feet a, b, and c, RPP values for MTB-3 (x-axis) and MTB-4 (y-axis) are shown. Regression lines are fitted to each set of points. The angular distance values $\Theta(\cdot)$ are small for the pairs of regression lines (a,b), while $\Theta(a, c)$ and $\Theta(b, c)$ are large, quantifying that the slope of the fit for c is different from the fit for a and b, respectively. MTB: metatarsal bone.

Two feet are similar if the slopes of most of the $\binom{8}{2}$ regression lines are similar, taking into account the goodness of fit of each line. Let i be a session, and let s, t be two plantar regions. The regression line's slope between s and t for session i is denoted as $m(i, s, t)$ and the Pearson correlation coefficient as $cor(i, s, t)$, quantifying the goodness of fit of the regression lines. The dissimilarity between two sessions i, j for these two regions is then defined as the difference between the slopes of the regression lines $m(i, s, t)$ and $m(j, s, t)$, given by

$$\Theta(i, j, s, t) = \tan^{-1}\left(\frac{m(i, s, t) - m(j, s, t)}{1 + m(i, s, t) \cdot m(j, s, t)}\right). \tag{7.3}$$

We then sum the angular distances over all pairs of plantar regions to obtain a value expressing the distance between two feet $d_{\text{pairs}}(i, j)$, defined as

$$d_{\text{pairs}} = 1 - \sum_r \sum_s \Theta(i, j, s, t) \cdot \frac{|cor(i, s, t)| + |cor(j, s, t)|}{2} \tag{7.4}$$

(if $i \neq j$; otherwise 0), where $cor(i, s, t)$ is the Pearson correlation coefficient between the plantar pressure records of r and s for foot i. Because the angular distance depends on the goodness of fit of the regression, angular distances of more reliable regression lines were assigned a higher weight in the distance calculation.

Similarity based on centers of pressure. For the third variant of similarity, we cluster the average RPPs in each region. Two feet are similar when most of the k centroids are similar. For each region, we call k-means multiple times with different k and select the run with the optimal Silhouette coefficient. However, due to a possibly different number of clusters for the different regions, we consider that the distance between two centroids may be smaller for a clustering with many clusters than for a clustering with only a few clusters. Therefore, we add a weighting factor so that the $d_{\text{centers}}(i, j)$ is defined as:

$$d_{\text{centers}}(i,j) = \sum_{s} |ctr(i,s) - ctr(j,s)| \cdot \log(|C(s)|) / \sum_{t} \log(|C(t)|) \tag{7.5}$$

where $ctr(i, r)$ is the centroid of the cluster to which i has been assigned with respect to clustering for sensor s and $|C(s)|$ is the number of clusters for the clustering of s.

7.2.3 Identifying Foot Profiles by Clustering

To identify distinct pressure pattern subgroups, we carry out k-medoids [150], DBSCAN [72], and agglomerative hierarchical clustering.

k-medoids. Like k-means and X-means (recall Section 5.2.1), a k-medoids cluster is represented by an instance for which the sum of the distances to all observations of the same cluster is minimal. The difference between a k-means centroid and a k-medoids medoid is that the latter must be an actual observation. In contrast, a centroid is an artificial observation derived from mean values. Compared to k-means, k-medoids can be more robust to outliers [251] and has, at least for this application, an arguably more intuitive cluster representation: The medoid can be interpreted as the most representative foot in the cluster, whereas a centroid can be very different from the pressure distribution of any patient within the cluster. Thus, this notation allows us to use clustering to examine temporal patterns of plantar pressure by visualizing the cluster medoids' RPP time curve.

DBSCAN. DBSCAN [72] partitions instances into clusters based on their estimated density distribution and builds clusters of arbitrary shape and size. Pressure distributions of feet that differ considerably from any cluster are declared as outliers, which may be abnormal pressure patterns of patients potentially requiring medical supervision. DBSCAN has two parameters: the radius *eps* defining the "neighborhood" of an instance and the minimum number $minPts$ of neighbors for an instance to be a core point (recall Section 6.3.1).

Hierarchical clustering. In agglomerative hierarchical clustering, similar instances are iteratively merged into clusters in a bottom-up fashion. The order in which two clusters are merged depends on the linkage strategy: In *complete linkage*, the distance between two clusters is defined as the maximum distance between a pair of instances (one from each cluster). The two clusters that minimize this maximum distance are selected for merging. Using a dendrogram, it is possible to break down the "tree" of clusters to understand the progressive merging process. Optionally, a parameter k can be specified to obtain a specific partitioning with k clusters.

7.2.4 Evaluation Setup

Preprocessing. Some of the sensor recordings were identified as noisy or erroneous. To reduce the impact of such *extreme* recordings, these values were replaced with the median of all recordings for the sensor in that session. Following the *inner fence* outlier definition in boxplots [132], a value x was flagged as extreme if it fell outside the range $[Q_1 - 1.5 \cdot (Q_3 - Q_1), Q_3 + 1.5 \cdot (Q_3 - Q_1)]$, where Q_i is the i^{th} quartile and $(Q_3 - Q_1)$ is the interquartile range over all sensors in i. Then the time curves were smoothed using locally weighted scatterplot smoothing (LOWESS) [52] with a smoother span of 5%. Recordings from all stance episodes were used for distance calculation and clustering.

Experimental setup. For part A, we use the three distance measures from Section 7.2.2 to cluster the participant feet, using the algorithms k-medoids, DBSCAN, and hierarchical agglomerative clustering (Section 7.2.3). For DBSCAN, we estimate an appropriate value of *eps* using the elbow method presented in Section 6.3.1, where k is set to $minPts$. For hierarchical clustering, we cut the dendrogram to obtain k clusters. Cluster quality was measured by the Silhouette coefficient (recall Section 4.2.1). For k-medoids, we vary k from 2 to 10. For DBSCAN, we set $minPts$ from 2 to 10 and automatically determine *eps*. For hierarchical clustering, we set the number of clusters from 2 to 10 and use single linkage, complete linkage, and average linkage, respectively. For part B, we restrict to the best performing combination of distance measure and clustering algorithm of part A.

Included datasets. In part A, foot insole sensor recordings of 20 patients (5 female, 15 male, 66.2 ± 8.4 years) with type 1 or type 2 diabetes and sensomotoric peripheral polyneuropathy were available. 19 participants performed the experiment twice, the remaining participant only once. For 34 of these 39 experiments, recordings of both feet were available, reaching a total of 73 experiment datasets.

For part B, data for 25[1] diabetes patients (6 females, 19 males, age 64.8 ± 9.8 years) and 18 non-diabetic healthy volunteers (10 females, 8 males, age 62.9 ± 7.6 years) were available. For all 43 sessions, both feet' sensor recordings were available, which were used independently, reaching a total of 86 session datasets.

7.3 Validation on DIAB

7.3.1 Results on Part A

Table 7.1 shows the clustering results. The maximum Silhouette score is achieved by k-medoids with d_{RPP} (k = 4; Silh = 0.78). The number of clusters found by the 3 clustering algorithms differs: while k-medoids finds 4 clusters for each distance measure, DBSCAN returns one large cluster and some outliers (i.e., noise points); for hierarchical clustering, the number of clusters varies the most, from 2 to 10. k-medoids outperforms DBSCAN and agglomerative hierarchical clustering for all three distance measures. Besides, d_{RPP} outperforms all other distance measures with respect to the three algorithms.

For the best clustering (d_{RPP}, k-medoids), a boxplot-based visualization of the RPP distributions of the medoids for each sensor is shown in Figure 7.3. While the RPP distribution of medoid

[1]The data for this chapter became available only gradually. Data analysis for part B was conducted after part A. At that time, additional data of diabetes patients were available. We only considered data from patients for whom recordings for both feet were available.

Table 7.1: **Clustering results (part A).** Optimal number of clusters k_{opt} and Silhouette coefficient $Silh_{opt}$ for each distance measure and algorithm.

Algorithm	k_{opt}	$Silh_{opt}$
$d_{\mathbf{RPP}}$		
k-medoids	4	0.78
DBSCAN	1 + 14 noise points	0.57
Hierarchical (single linkage)	2	0.40
$d_{\mathbf{pairs}}$		
k-medoids	4	0.31
DBSCAN	1 + 2 noise points	0.13
Hierarchical (average linkage)	3	0.18
$d_{\mathbf{centers}}$		
k-medoids	4	0.45
DBSCAN	1	0.12
Hierarchical (average linkage)	10	0.18

1 is balanced, medoid 2 has a pressure focus on the centrally located MTB-3; medoid 3 has an overall higher pressure load, especially for Digitus-1, MTB-2, MTB-3, MTB-5, and Lateral, while medoid 4 represents a "skewed" pattern with increasing pressure from the medial to the lateral forefoot (as measured at MTB-5). The substantial differences in the RPP distribution of these medoids suggest high variability in how DFS patients apply pressure to their feet. Determining patient subpopulations serves as a basis for reducing the risk of DFS-related foot complications by supporting early detection, treatment, and prevention [19, 287, 39, 60, 173, 43], for example, by personalized footwear tailored to patient needs [23].

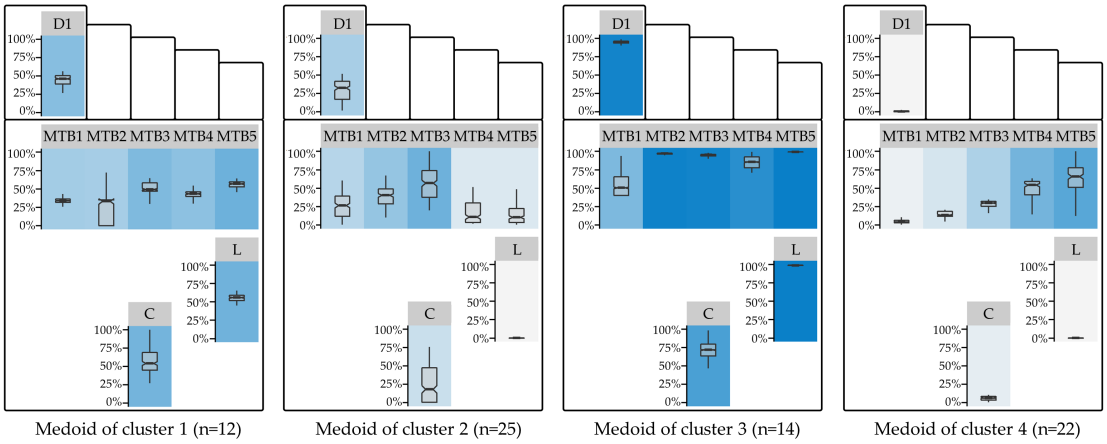


Figure 7.3: **Cluster medoids of the best clustering.** Visualization of the RPP distribution of the medoids of the best clustering (d_{RPP}; k = 4). The boxplot panel background color represents the median RPP, from light gray (low median RPP) to dark blue (high median RPP). D1: Digitus-1; MTB: metatarsal bone; C: Calcaneus.

The optimal clustering resulted in 4 subgroups. This number is consistent with the results of Deschamps et al. [60] (for the diabetes group) and De Cock et al. [59]. Furthermore, some of our findings are similar to the study of Bennetts et al. [23]. The RPP distribution of the second cluster's medoid corresponds to the "central pattern" of De Cock et al. [59] and cluster 2 of Deschamps et al. [60]. Being the medoid of the largest of the four clusters (n = 25 out of 73

feet (ca. 34%)), it is characterized by low to medium RPP the with focal point of pressure on the central forefoot regions MTB-3. Our fourth medoid is similar to cluster 4 of Deschamps et al. [60]. However, while cluster 4 is the second largest in our study (n = 22 feet (ca. 30%)), the relative proportion in Deschamps et al. [60] is much smaller (n = 30 of 194 feet (ca. 15%)). While RPP pressure is rather low in the medial regions, median RPP gradually increases toward the lateral forefoot regions. The balanced, moderate pressure on all regions of the first medoid (n = 12 (ca. 16%)) represents a well-distributed pressure loading pattern. It is similar to the largest cluster (cluster 4) of Bennets et al. [23]. The overall high pressure loading of the third medoid may be an indicator of an adverse posture. Patients from cluster 3 (n = 14 (ca. 19%)) may be overloading their feet and should therefore be warned more urgently than the other subgroups.

7.3.2 Results on Part B

Figure 7.4 shows the Silhouette coefficient and the optimal number of clusters for each subgroup. Patients, controls, and the combination of both groups can be best described with four plantar pressure profiles. For the clustering with the optimum k of each group, a summary of the relative plantar pressure distribution for the session datasets of each cluster is provided in Figure 7.5.

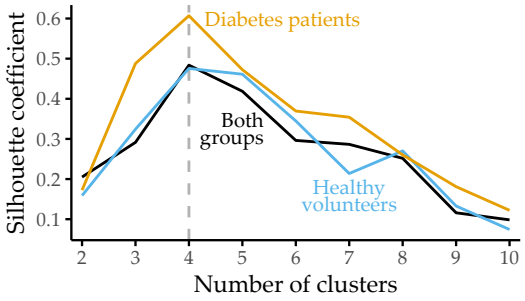


Figure 7.4: **Silhouette coefficient for different numbers of clusters.** Silhouette coefficient for each group and k-medoids clustering using the distribution of eight plantar pressure regions with the number of clusters k set between 2 and 10. For each group, the best clustering is obtained with k = 4 clusters.

Healthy volunteers. For cluster 1, representing 50% of the volunteer participants, the median RPP is high in all regions. Clusters 2 and 3 show some variability for pressure in the forefoot regions. Cluster 4, characterized by low median RPP at MTB-1 and MTB-5, describes only 2 of 36 feet in the volunteer group.

Diabetes patients. Cluster 1 is the largest subgroup and is characterized by high plantar pressure (median RPP above 80%). Cluster 2 represents an evenly balanced median RPP profile. Here, median relative plantar pressures for Digitus-1, MTB-1, and MTB-5 range from 30% to 50% of maximum, while median relative plantar pressures for the central forefoot (MTB-2, MTB-3, MTB-4), Lateral, and Calcaneus were recorded between 50% and 75%. For cluster 3, the median RPP is above 80% of all regions' maximum values except for the medial regions Digitus-1 and MTB-1. The latter cluster has the highest variance of all four clusters, with a high difference between the first and third quartiles for several regions. For cluster 4, high median relative plantar pressures are perceived at all MTB sensors (RPP $> 75\%$) and almost no pressure on Lateral (median RPP $< 2\%$).

Table 7.2: **Baseline characteristics.** Characteristics of diabetes patients (P) and healthy volunteers (V) with respect to the clustering on the combined group.

	Cluster 1		**Cluster 2**		**Cluster 3**		**Cluster 4**	
Characteristic	**P**	**V**	**P**	**V**	**P**	**V**	**P**	**V**
n	9	19	9	13	10	4	8	0
Sex (f/m)	2/7	9/10	4/5	8/5	4/6	3/1	1/7	–
Age (years)	61 ± 9	64 ± 7	68 ± 5	60 ± 9	68 ± 8	67 ± 6	63 ± 10	–
Height (cm)	178 ± 5	172 ± 8	173 ± 5	169 ± 12	173 ± 6	166 ± 11	178 ± 6	–
Weight (kg)	97 ± 22	79 ± 13	85 ± 15	76 ± 11	92 ± 15	69 ± 18	100 ± 18	–
BMI	31 ± 6	27 ± 3	28 ± 4	27 ± 4	31 ± 5	25 ± 6	31 ± 4	–

When we perform clustering on the combination of both groups, some clusters of the diabetes patients merge with the ones of controls. The optimal number of clusters to describe the pressure distribution best remains 4. Clusters 1-3 contain both patients and controls, but cluster 4 summarizes pressure distribution patterns only found in diabetes patients with severe polyneuropathy.

Table 7.2 summarizes the data for diabetes patients and healthy volunteers within the combined group's four clusters. In clusters 1-3, the mean values of weight, height, and BMI for patients are higher than those for volunteers. Since cluster 4 has the highest values for height, weight, and BMI, it can be assumed that these anthropometric characteristics also influence the measured pressure values and that the class separation cannot be explained exclusively by pathology.

7.4 Discussion of the Findings from the Medical Perspective

Four pressure clusters each were identified for healthy volunteers and diabetes patients. These clusters reflect the way the groups distribute pressure to foot regions vulnerable for ulceration. The ultimate goal of these profiles is to reduce ulcer risk: while clinical examination remains essential, pressure distribution analysis, as proposed here, can serve as the basis for preventive strategies [19, 287, 39, 60, 57, 173, 43], including the manufacturing of footwear tailored to patient needs [23]. An interesting finding is that an unsupervised procedure alone, i.e., without using the target variable (presence of diabetes), generated a cluster consisting exclusively of diabetes patients, see Figure 7.5 (c).

This result is remarkable because the study protocol, with its strict sequence of standing and sitting episodes, specifies exactly when and for how long pressure loads are to be applied. Furthermore, during the experiment, participants were instructed to adhere to the protocol, i.e., to apply pressure continuously during the standing episodes. It is possible that the diabetes patients adhered more strictly to the study protocol due to their neuropathy. It can also be assumed that a pain-related fatigue effect occurs more frequently during the second loading phase over 20 min (stance episode 6) compared to the first (stance episode 3), indicated by a larger percentage of recordings of pressure release. To quantify these changes, we calculated the percentage of recordings in which the participants released pressure during stance episodes 3 and 6 (Figure 7.6). No significant intra-group differences between stance episodes 3 and 6 were found. However, the mean percentage of pressure release recordings was significantly different ($p < 0.01$) between healthy volunteers and diabetes patients with sensoric polyneuropathy.

Comparing the clusters for healthy subjects and diabetes patients revealed a high overlap in

the plantar pressure distribution. For example, cluster 1 and cluster 3 of healthy subjects and diabetes patients are characterized by almost identical pressure loading patterns. To investigate whether a group-specific pressure distribution could be identified, a cluster analysis was performed for the combination of both groups. A random sample of diabetes patients was selected to ensure a balanced distribution between the two groups. The observed pressure patterns in Figure 7.5 (c) support the assumption of common pressure patterns: clusters 1, 2, and 3 comprise the pressure distributions of both groups. However, cluster 4 was unique to diabetes patients. This pattern is characterized by the lowest median relative plantar pressure on the Lateral and Calcaneus.

Also, unlike other studies, we avoided k-means because it is sensitive to outliers and because cluster centroids are composite objects, not true feet; rather, we provide a cluster's medoid, which is the most representative foot in the cluster, while a cluster centroid (under k-means) is a derived vector of averages that may be very different from any patient's pressure distribution inside the cluster.

Some of the observed differences compared to related studies could be because of differences in study participants, population sizes, experimental protocols, and measurement devices [59, 23, 60]. For example, neither the medial M1 pattern nor the M2 pattern of [59], which focuses on either MTB-1 or MTB-2, could be replicated. Besides, some of the results were not observed in previous studies, such as the pattern disparity between lateral and other regions in cluster 4 (diabetes patients).

Our main finding is that a group of patients (cluster 4) applies plantar pressure in a way that is not found in healthy volunteers. To draw conclusions from the other clusters, it is necessary to consider a larger sample of healthy individuals, which would allow reducing idiosyncrasies possibly found here and form homogeneous clusters of healthy individuals.

The study protocol's simplicity makes it easily reproducible, but the protocol cannot capture the complexity of foot movement in everyday life. Thus, a limitation of the study is to only detect differences in pressure distribution that can be detected by changes in posture. When patients apply pressure in an uncontrolled environment, the variance between their pressure profiles will inevitably increase. Therefore, a study that considers more effortful activities of the patients (running, cycling, climbing stairs, etc.) is needed to complement the results.

Another limitation of the results is the small sample size, which prevented examining differences between female and male participants because the gender samples would have been too small for generalization.

Finally, a long-term follow-up would allow us to determine which patients eventually develop ulceration. This would provide a critical endpoint and evaluate whether certain clusters are associated with a higher risk of developing diabetic foot syndrome.

7.5 Conclusion

We have proposed three similarity measures for short temporal sequences extracted from raw mHealth data to create representations of distinct subpopulations via clustering. The DIAB data served for proof-of-concept validation. The similarity measures were based on (i) relative plantar pressure, (ii) pressure distribution in pairs of sensors, and (iii) centers of pressure. We found four distinct subpopulations of plantar pressure patterns.

Our approach needs to be validated on larger cohorts. The modeling of similarity can be transferred from pressure to temperature recordings in the plantar regions, possibly closing the gap in early detection of tissue damage and ulcer formation. The formation of a diabetes

patient cluster emphasizes pathological differences, i.e., impaired sensation and microcirculatory defects leading to tissue damage. Our results may lay the groundwork for approaches to identify pressure (and temperature) patterns predictive for emerging foot problems with diabetes in the future.

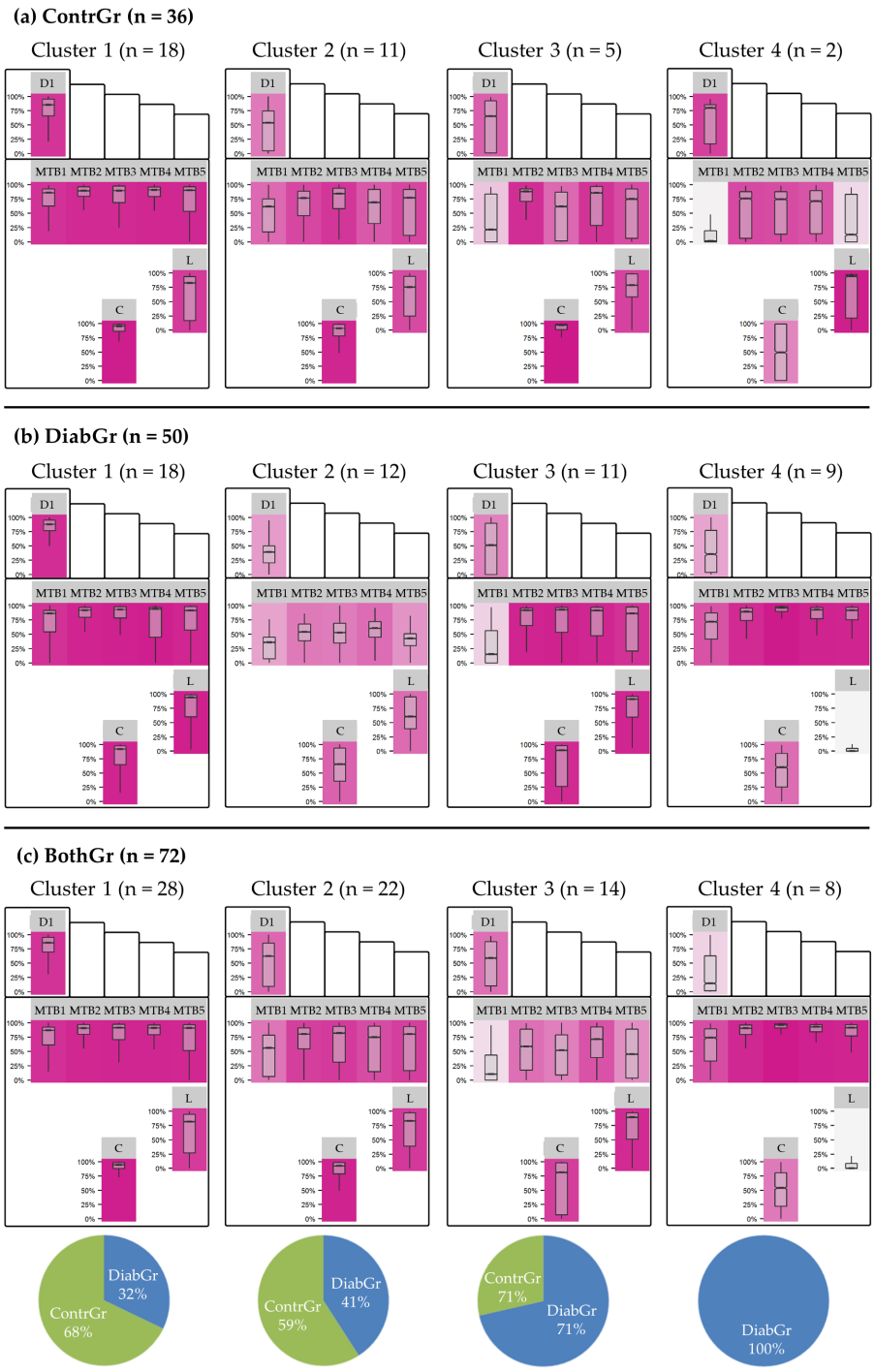


Figure 7.5: **Summary visualization of intra-cluster RPP distributions.** Relative plantar pressure distribution for each cluster and sensor over the whole session (including pauses). Panel background of boxplots depicts median relative plantar pressure with a linear color gradient, from light gray (low relative plantar pressure) to violet (high relative plantar pressure). Pie charts show the percentages of healthy volunteers and diabetes patients for each cluster in the combined group. D1: Digitus-1; MTB: metatarsal bone; C: Calcaneus. The figure is adapted from [200].

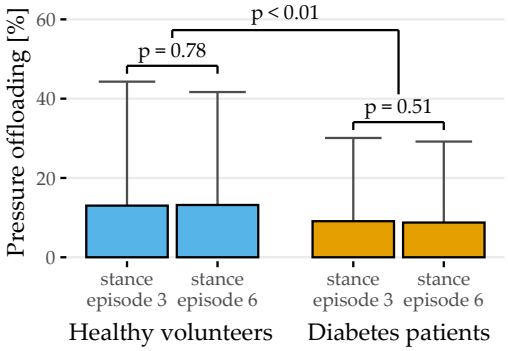


Figure 7.6: **Average percentage of recordings without pressure application during 20-minute standing episodes.** Intermittent pressure release in diabetes patients with neuropathy occurs less frequently than in healthy controls. The underlying dataset is from our previous publication [202] where 114 propensity-score matched observations (57 from volunteers, 57 from diabetes patients) are analyzed. For intra- and inter-group statistical comparison, a two-sided Student's t-test is used. The figure is adapted from [202].

Part III

POST-MINING FOR INTERPRETATION

Chapter 8

Post-Hoc Interpretation of Classification Models

Brief Chapter Summary

Complex *black-box* models can have high predictive performance but are difficult to interpret. We propose an end-to-end data analysis workflow for high-dimensional medical data that includes steps for data augmentation, modeling, interleaving model training with feature elimination, and post-hoc analysis of the trained models. Our approach delivers statistics and visualizations representing global feature importance, instance-individual feature importance, and subpopulation-specific feature importance. We validate our workflow on three modeling tasks: (i) tinnitus-related distress after treatment in tinnitus patients, (ii) depression at baseline 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: *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 socio-demographics". In: *Scientific Reports* 10 (2020), pp. 1-9. 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 [265]. Obtaining results that both data scientists

and medical experts easily understand 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 followup 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 non-linear, it is often difficult to explain them understandably. Arising from the need to provide understandable insights into otherwise opaque models, the *interpretability* community of machine learning has gained traction, intending to resolve the dilemma of choosing between moderately accurate but interpretable models and highly accurate but opaque black-box models [196, 2, 42]. This chapter describes an end-to-end data analysis workflow for high-dimensional medical data that includes steps for data augmentation, modeling, interleaving model training with feature elimination, and post-hoc analysis of the trained models. We validate our approach on three datasets.

This chapter is organized as follows. Section 8.1 describes reasons for using interpretable machine learning methods and provides methodological underpinnings of selected pioneering methods. Subsequently, we present the components of our mining workflow in Section 8.2. In Section 8.3, we validate our workflow on three datasets. Section 8.4 discusses our findings from the medical perspective. Section 8.5 concludes the chapter.

8.1 Motivation and Methodological Underpinnings

Current state-of-the-art machine learning algorithms, such as gradient boosting [85] for tabular data and deep learning [100] for unstructured data (images, videos, audio recordings), are designed 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, medical experts may face the dilemma of choosing either an opaque black-box model with high predictive power or a simple, less accurate model that can be explained to the domain expert. Especially in high-risk domains such as healthcare, where misconceptions can have serious consequences, the ability to explain a model's reasoning is a highly desirable (sometimes mandatory) property of any decision-support system [105, 196].

As a result, methods that explain the predictions of complex machine learning models have attracted increasing attention in recent years [42, 2]. Existing methods are classified according to different criteria [196]. 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 [151]. 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 the root to leaf nodes provide insights into feature interactions. Moreover, 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 cannot capture linear, non-axis-parallel relationships between predictors and response, and they can be unstable for small value changes in training data [118]. Therefore, they may be unsuitable for complex learning tasks.

If a more sophisticated model is trained, post-hoc methods can be applied to examine the model after training. These methods' outputs can be *feature summary statistics*, *model internals*, *individual observations*, and *feature summary visualizations* [196]. In general, feature summary statistics are individual scores that express the overall importance of a feature to the 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 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.

8.1.1 Selected Model Interpretation Methods

We discuss partial dependence plots [85] (PDP), Local Interpretable Model-Agnostic Explanations [226] (LIME), and Shapley Additive Explanations [183] (SHAP), with emphasis on the interpretation of their outputs.

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* [85] (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 predictor and the response estimated by the model.

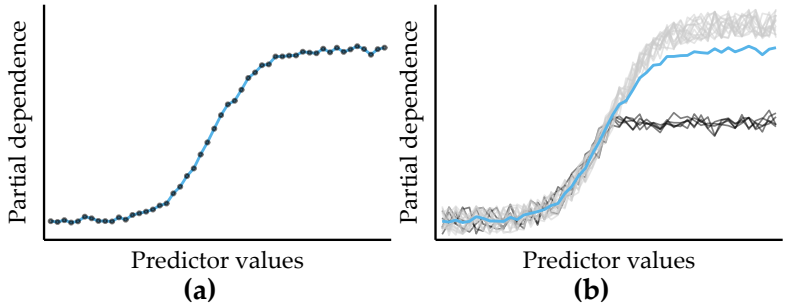


Figure 8.1: **Illustrations of a partial dependence plot (PDP) and an individual conditional expectation (ICE) plot on artificial data.** (a) PDP for a predictor on an artificial dataset. Points represent a sample of the predictor distribution. (b) PDP augmented with ICE curves. There are two distinct 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. According to Friedman [85], 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 the dataset 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. [99] proposed *individual conditional expectation* (ICE) plots to show a curve for each observation. Figure 8.1 (b) illustrates an example of 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, only one or a small number of selected observations. *Local Interpretable Model-Agnostic Explanations* [226] (LIME) is a popular local post-hoc interpretation method. LIME's central assumption is that a complex model is linear on a local scale [226]. 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, where Figure 8.2 (a) shows the decision boundary of a black-box model. Since the non-linear decision boundary is quite complex, the model and its predictions cannot be explained in simple terms. LIME attempts to approximate the black-box model's behavior by creating a linear surrogate model that performs particularly well in the vicinity of a user-selected instance of interest. To this end, a *perturbed* training set is created by repeatedly randomly changing the instance of interest values. 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 horizontal bar length represents the model coefficient of a feature. While LIME provides intuitive interpretations and applies 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 method and number of features considered for the surrogate model. The stability of the results of LIME has been critically discussed [5, 279].

Model-specific interpretation methods are limited to specific model families, while *model-agnostic* interpretation methods can be applied to any model type. Model-specific methods are based on model internals and are widely used for neural networks [232], e.g., layered relevance propagation [15], which explicitly uses a neural network's layered structure 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 a model's output, 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.

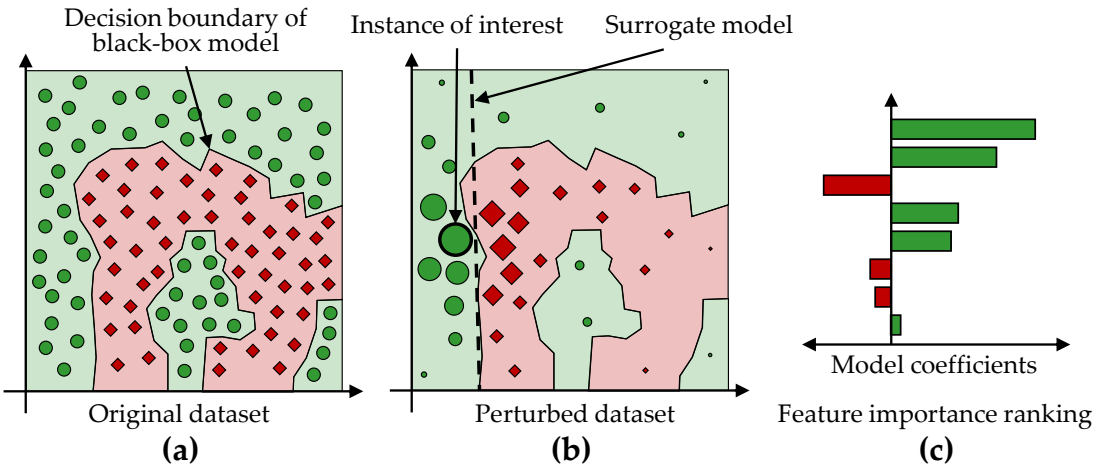


Figure 8.2: **Illustration of LIME's main ideas.** (a) A data set with a two-class problem represented as a two-dimensional scatterplot for simplicity. The non-linear decision boundary of a black-box model cannot be easily explained. (b) LIME aims to approximate a black-box model's predictions in the 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.

SHAP. Closely related to LIME is the *Shapley Additive Explanations* [183] (SHAP) framework, which derives additive feature attributions to a model's predictions. SHAP is based on *Shapley values* [179, 255, 245], originally developed for game theory. The term "additive" denotes that for a given observation, the model output should be equal to the sum of attributions over all features. 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:

- The sum of the feature attributions for an observation is equal to the difference between the model's average prediction and the actual prediction for that observation (*local accuracy*).
- 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*).
- 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 [183], which works on any model type, and TreeShap [184] for tree-based models.

Feature Selection. In Section 5.1, we discuss dimensionality reduction techniques and conclude that these methods can only be used to a limited extent since the original dimensions' semantics are lost during the projection. An interpretation of the transformed data space in the application context is thus hardly possible. Feature selection (FS) can help with this limitation. In the context of predictive modeling, FS methods reduce the number of predictors 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 [31] and neural networks [275, 100]. Others, such as linear and logistic regression models, are susceptible to correlated predictors. Often domain experts require intrinsically interpretable models [265], which requires eliminating predictors that do not contribute substantially to the model performance.

Traditionally, FS methods are broadly classified into three categories: embedded, filter, and wrapper [108]. 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 [221, 162], regularization methods such as *least absolute shrinkage and selection operator* [84] (LASSO), and Ridge [134] regression. Filtering methods rank predictors only once based on some measure of importance, e.g., correlation with the target feature. Popular examples include correlation-based feature selection [112] (cf. the application in Section 6.3.3) and Relief [154]. 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 [46].

8.2 Overview of the Mining Workflow

In this section, we describe the four components of our mining workflow, which are:

1. Data augmentation, including the construction of new features and correlation analysis between features (Section 8.2.1),
2. Modeling of the learning task (Section 8.2.2),
3. Interleaving model training and feature elimination (Section 8.2.3), and
4. Post-hoc analysis of the learned models (Section 8.2.4).

8.2.1 Data Augmentation

To increase model performance, manual derivation of predictive features (feature engineering) is required in many applications (recall the evolution features presented in Chapter 6). For example, with image data, a general approach is first to derive descriptive features and transform the data into a tabular format to use off-the-shelve classifiers. Therefore, we derive predictive features from image data in the first step of our workflow. We also explore correlations between the (derived) predictors as a step of exploratory data analysis.

8.2.2 Modeling of the Learning Task

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

- Least absolute shrinkage and selection operator [84] (*LASSO*) and *Ridge* [134] are extensions of ordinary least squares (OLS) regression that perform feature selection and regularization 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 coefficients' absolute values, 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 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 that capture as much 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 [65].
- A support vector machine (*SVM*) [31] learns linear or non-linear 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*, i.e., the separating hyperplane with the maximum margin to the support vectors. In case a linear decision boundary does not exist, non-linear 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 artificial neural network (*NNET*) consists of a structure of nodes connected 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 to which it forwards values. 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 can extract more classification-relevant feature sets compared to expert-driven feature engineering or 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 [100]. We used a feed-forward NNET with one intermediary layer (*hidden unit*) [275].
- Weighted k-nearest neighbor [120] (*WKNN*) is a variant of KNN classification. To classify an observation with an unknown response value, the k *nearest* training observations are

identified, and the modus of their response values are used as the prediction. The 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 considers 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 obtain more reliable class-conditional probability estimates.
- Classification and regression trees [34] (CART), C5.0 [221], random forests [35] (RF) and gradient boosted trees (GBT) [85] are tree-based models. Algorithms 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 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 gradient boosted model, the base trees are constructed and added to the composite model so that any new tree reduces the error of the current set of trees.

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 the 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 8.1). 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 the 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 8.1: **Overview of hyperparameter tuning grid.** All classifiers were implemented with the statistical programming language R [261] using the package `mlr` [29], 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` [84])	`lambda`	10^{-2}	10^{10}	100
GPLS (`caret` [161])	`ncomp`	1	5	5
SVM (`e1071` [194])	`cost`	0.01	3	6
	`gamma`	0	3	4
	`kernel`	–	–	4*

Algorithm (R package)	Parameter	Min.	Max.	No. of values
NNET (`nnet` [275])	`size`	1	13	7
	`decay`	10^{-4}	1	6
WKNN (`kknn` [120])	`k`	1	77	20
NB (`e1071` [194])	`laplace`	1	5	5
CART (`rpart` [263])	`cp`	0.001	0.1	5
C5.0 (`C50` [163])	`CF`	0	0.35	7
	`winnow`	`FALSE`	`TRUE`	2
	`rules`	`FALSE`	`TRUE`	2
RF (`ranger` [296])	`mtry`	4	100	7
	`min.node.size`	1	25	6
GBT (`xgboost` [49])	`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

8.2.3 Iterative Feature Elimination

We employ a feature selection wrapper that successively eliminates a subset of predictors that do not *positively* contribute to a model's performance. A predictor's contribution is computed using *model reliance* [79], which is a generalization of random forest permutation feature importance [35]. 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. Shuffling a predictor's values removes any relationship between the predictor and the target variable. Hence, it is assumed that permuting a more important predictor leads to a higher decrease in accuracy as opposed to a predictor with a lower model contribution.

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 the 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 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.

8.2.4 Post-Hoc Interpretation

SHAP. To facilitate model interpretation, we use the model-agnostic post-hoc framework SHAP [183, 184] 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 the 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 the 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 are calculated for the best model ζ_{opt} according to AUC. A ranking of each feature's attribution towards ζ_{opt} is 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 is clustered with agglomerative hierarchical clustering to identify subgroups of patients with similar SHAP values.

PDP feature importance. We derive a *global* feature importance measure from the PD of a predictor. We assume that predictors with high PD variability are important. Consider the two PD curves in Figure 8.3: the predicted response changes considerably with different predictor values for the blue PD curve, whereas the green PD curve is 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 the 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.

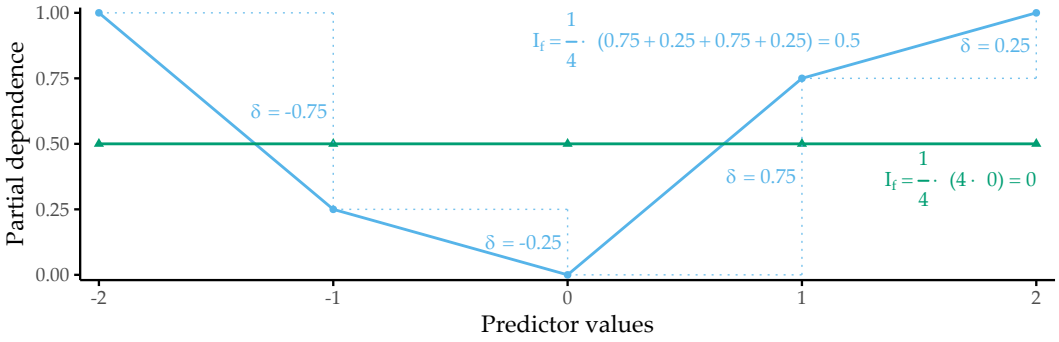


Figure 8.3: **Illustration of partial dependence importance.** Partial dependence importance I_f for two exemplary PD curves.

8.3 Validation on Three Datasets

We validate our workflow on three datasets:

- *CHA-Tinnitus*: CHA data (recall Section 2.2.2) with tinnitus-related distress at baseline (T0) as target variable,
- *CHA-Depression*: CHA data with depression after treatment (T1) as target variable, and
- *ANEUR*: ANEUR data (Section 2.2.4) with rupture status as the target variable.

Figure 8.4 illustrates our workflow adapted to the datasets.

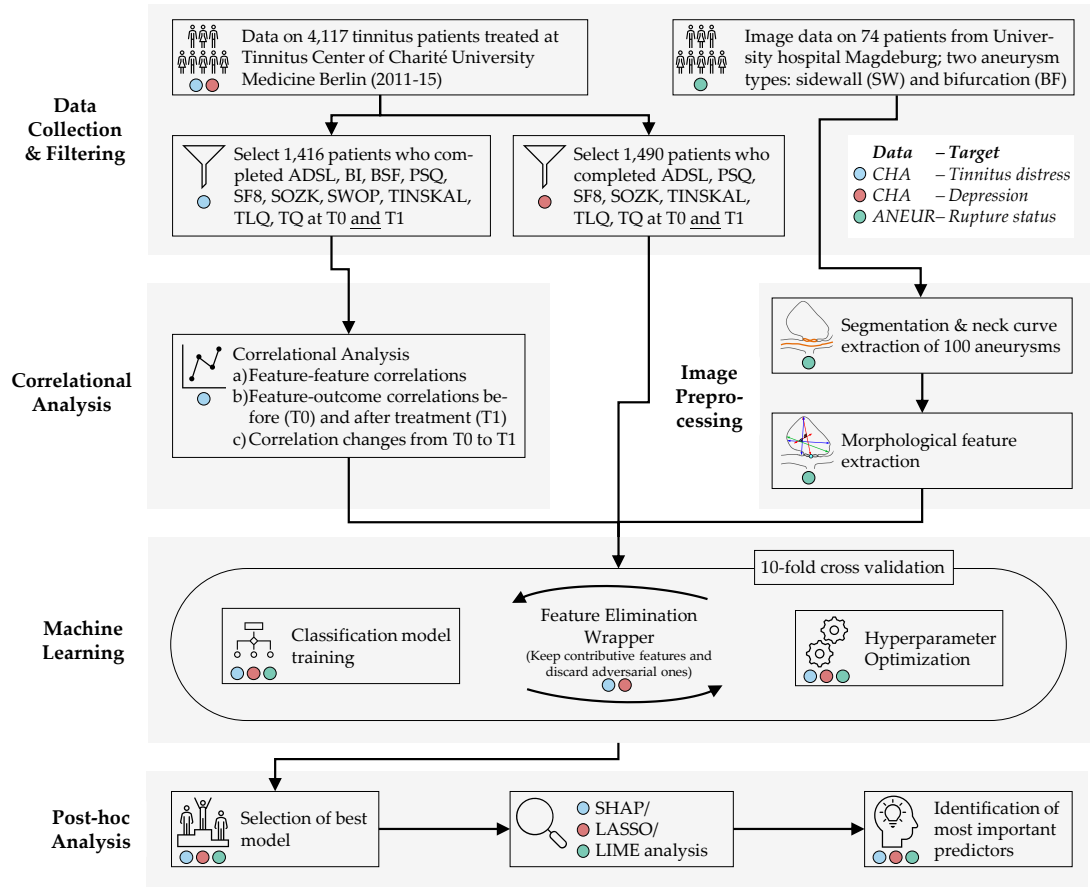


Figure 8.4: **The (dataset-specific) mining workflow.** For the CHA dataset, we select patients with complete data for each of the two classification tasks. For ANEUR, we segment the aneurysms from the raw image data, perform automated centerline and neck curve extraction, and generat the morphological features. We perform 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 embedd model training in an iterative feature elimination wrapper that retains predictors identified as important to the model. We select the best overall model based on AUC and used post-hoc interpretation methods to identify predictors with the highest attribution to the model prediction on a global, subpopulation and observation level.

Hereafter, we report our results on all three learning tasks, i.e., regarding CHA-Tinnitus (Section 8.3.1), CHA-Depression (Section 8.3.2) and ANEUR (Section 8.3.3).

8.3.1 Results for CHA-Tinnitus

The learning problem. We use baseline (T0) data to predict the tinnitus patients' distress at the end of the therapy (T1). We derive our target variable TQ_distress from the total score of "the Tinnitus Questionnaire" [98]. The mean tinnitus-related distress score decreased between T0 to T1 from 38.3 ± 17.1 to 31.7 ± 17.2, indicating a positive effect of the multimodal treatment. We apply the cutoff at 46 suggested by Goebel and Hiller [98] to distinguish between patients with *compensated* or *decompensated* tinnitus.

Correlational analysis. We calculate the Spearman correlation coefficient between each pair of predictors. Using agglomerative hierarchical clustering with complete linkage, we arrange predictors in a correlation heat map visually identify predictor subgroups with similar intra-group and inter-group correlations. 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. Also, we identify predictors with the highest absolute correlation with the response at T0 and T1, respectively. Finally, we examine predictors whose correlation values with the TQ distress score differ most between T0 and T1. Figure 8.5 (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 8.5 (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 8.5 (a)) contains 47 predictors (ca. 22.9%) with positive wording, e.g., the SF8 mental health score (SF8_mentalhealth) and the BSF elevated mood score (BSF_elevatedmood). Predictors of one of the two subgroups exhibit a moderate to high negative correlation with the other subgroup predictors. Figure 8.5 (b) compares the correlation of the predictors with TQ_distress before (x-axis) and after treatment (y-axis). Overall, only low to moderate bivariate correlations are observed, as all values are between -0.6 and +0.6. The average absolute correlation change between T0 and T1 is 0.031. The change in absolute correlation is smaller than 0.067 for ca. 95% of the predictors (compare the distance of the points to the diagonal line in Figure 8.5 (b)). For 137 out of 205 predictors (66.8%), absolute correlation decreased from T0 to T1. Median target-correlations 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 8.5 (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 the largest correlation magnitude before ($\rho = 0.630$) and after treatment ($\rho = 0.564$). Figure 8.5 (e) shows the ten 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 8.6. 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 achieves 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 result 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 generates at least one model with

similar or even improved performance compared to the respective model learned on the whole set of predictors. In fact, except for WKNN, all classification methods benefit from feature elimination as they produce their best model on a predictor subset (cf. Figure 8.6). 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 tradeoff between high predictive performance and low model complexity, and therefore, we decide to investigate this model further.

Feature importance. For the best model, the attributions of the 26 selected features are shown in Figure 8.7 (a). Among the 26 features are 6 compound 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 the highest model attribution as it exhibits the highest average absolute SHAP value (change in log odds) of 0.448. The ADSL depression 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 is selected. Figure 8.7 (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 high range of 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, unlike the somewhat more dense bulk of points representing patients with SHAP values between ca. -0.7 and -0.4. This could indicate that patients who report high tinnitus impairment are more challenging 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 8.8 (1), which juxtaposes the predictor's actual values with the corresponding SHAP values for all patients and reveals a J-shaped relationship between them. More specifically, the predicted tinnitus-related distress decreases from 0 to 2.5, remains at a plateau from 2.5 to 4 and increases from 4 to its maximum value of 10. Besides TINSKAL_impairment, the features ADSL_depression, TINSKAL_loudness, BI_overallcomplaints, BSF_timestamp, and SWOP_pessimism also show a non-linear relationship with respect to their SHAP values.

Even though several predictors exhibit only a low to moderate global importance, some 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 patients with different education levels have an average SHAP value of -0.1 and hence are more close to the population average (cf. Figure 8.7 (b), Figure 8.8 (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 8.7 (b), Figure 8.8 (14)).

To investigate whether there are subgroups of patients with similar model explanations, we cluster the patients based on their SHAP values. Figure 8.7 (c) shows stacked patient-individual SHAP values for the six predictors with the highest average absolute SHAP values and the remaining predictors combined. According to the Bayesian information criterion (cf. insetted plot in Figure 8.7 (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 8.7 (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, depression, anxiety, tinnitus loudness, sleeplessness, pessimism, psychosomatic complaints, and 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. Also, cluster 1 contains a higher fraction of patients with lower secondary education ("Hauptschule"), report more frequently occurring 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 the highest physical health and levels of determination. Cluster 4 is somewhat close to the prediction average, which means that positive and negative SHAP values nearly even out. Concerning the average patient-sum of SHAP values, cluster 3 lies in between cluster 2 and cluster 4.

8.3.2 Results for CHA-Depression

The learning problem. We use baseline (T0) data to classify the tinnitus patients' depression severity. We derive our target variable from the total score of the *General Depression Scale* [222, 119] (ADSL). The mean depression score at T0 is 18.2 ± 11.7. Our target variable "depression status" is the dichotomized ADSL score. Following the recommendation of Hautzinger and Bailer [119], we distinguish between patients with *subclinical* (0-15) or *clinical* (16-60) depression.

Predictive performance of classification models. Figure 8.9 depicts the performance of all classification methods across iterations. The LASSO classifier achieves 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).

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 8.10 shows the median model coefficient for these features across ten cross-validation folds. From the ADSL questionnaire alone, 16 single items are included in the final model, including indicators of depression (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 contributes the highest number of predictors to the model. From the tinnitus-distress-oriented TQ, five predictors are selected. Further, the model uses another five 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).

Table 8.2 provides a description for each predictor from Figure 8.10.

Effect of feature elimination on classification performance. All classifiers but SVM show high stability in performance on smaller feature subsets. From Figure 8.9, 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 benefit 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 the most in performance from feature removal since their respective maximum AUC is obtained on the smallest predictor subset, with a cardinality of 22 and 10, respectively.

Table 8.2: **Most important features of LASSO model.** Predictors with the 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

8.3.3 Results for ANEUR

The learning problem. We predict the rupture status for a total of 100 intracranial aneurysms with morphological parameters. We learn different models for the subset of sidewall aneurysms (SW; 9 of 24 ruptured), for the subset of bifurcation aneurysms (BF; 29 of 62 ruptured), and on a combined group (43 of 100 ruptured) with 14 additional samples that could not be determined to be either SW or BF.

For ANEUR, it is necessary to extract morphological features from raw image data first.

Segmentation and neck curve extraction. Aneurysms and vessels are segmented using a threshold-based approach [95] from digital subtraction data reconstructed from 3D rotational angiography images. Subsequently, the centerline of the vessel is extracted using the *Vascular Modeling Toolkit* (VMTK, vmtk.org) [6]. Subsequently, the plane separating the aneurysm from its parent vessel is determined using the automatic ostium detection of Saalfeld et al. [230].

Morphological feature extraction. For each 3D surface mesh, we obtain the neck curve, the dome point D, and the two base points B_1 and B_2. As described in [230], B_1 and B_2 are approximated as points on the centerline with the largest distance where the rays from B_1 and B_2 to D do not intersect the surface mesh. Figure 8.11 illustrates the extracted parameters, where H_{max}, W_{max}, H_{ortho}, W_{ortho}, and D_{max} (Figure 8.11 (a)) describe the aneurysm shape [62, 172]. The angle parameters α, β, and γ (Figure 8.11 (b)) are extracted based on B_1, B_2, and D, 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 derive an estimate for the surface area A_A and volume V_A of the aneurysm (Figure 8.11 (c)). We provide two variants for the surface area of the ostium, A_{O1}, and A_{O2} (Figure 8.11 (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 [230]. Therefore, A_{O2} is extracted as a parameter comparable to other studies that often use a cutting plane to determine the ostium. Our method achieves a local optimum for highly lobulated aneurysms 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 8.3 provides an overview of all extracted morphological features.

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: **Overview of morphological features extracted for ANEUR**.

#	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)

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.

Concerning 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 8.14 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 trends but different intercepts. In contrast, the ICE curves of WKNN (Figure 8.14 (c)) appear more jittery. The plots summarize some of the idiosyncrasies of the different model families. 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 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.

8.4 Interpretation of the Findings from the Medical Perspective

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 ANEUR (Section 8.4.3).

8.4.1 CHA-Depression

Machine learning is used to build predictive models of depression severity based on structured patient interviews [181, 153]. 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 carries over the depression status at T0 as prediction for T1; such as model yields 79% accuracy. Our models outperform this baseline, although they are likely to 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 a timely prediction of depression severity and selecting appropriate treatment with only a few questionnaire items.

Consistent with previous studies [168], we have found a strong association between tinnitus distress and depression severity. Furthermore, predictors measuring perceived stress and demands are significant contributors to depression in tinnitus patients [266]. The fact that predictors are 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 [292], high sensitivity in detecting depression was achieved 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."* [292], 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 have identified a strong relationship between non-German citizenship and depression severity (cf. Figure 8.10 and Table 8.2). Although some studies reported ethnic differences in depression [227, 290], this item's occurrence tends to suggest higher perceived social stress in patients of predominantly Turkish origin due to higher unemployment rates, larger families, and lower housing conditions in this demographic group. Because only 5.0% of the cohort population were non-German citizens, these results could also result from overfitting. Since the associated predictor in the first iteration of feature elimination has a model reliance score of less than 1.0, it is 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 improve by feature selection, i.e., the AUC in the second or later iteration is superior to the AUC in the first iteration (where all 205 predictors are used). For example, the two decision tree variants achieve the highest performance on the smallest feature subset in each case. Regarding the LASSO classifier, which performs best, it is encouraging that only 6 predictors from 4 questionnaires show 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 socio-demographic predictors are 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. Costs can be measured not only by the financial expense of an examination, but also by the psychological or physical burden for a subject undergoing an examination (e.g., a painful biopsy vs. a blood test). For example, Yu et al. [298] performed feature selection under a budget, where the cost of feature acquisition was derived from medical experts' suggestions based on the total financial burden, patient privacy, and patient inconvenience. Kachuee et al. [149] derived 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 socio-demographic 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 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 correctly classifying 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 is removed in an early iteration could lead to a better model in a later iteration. A possible solution would be to backtrack or revisit earlier iterations if it turns out that some of the removed predictors actually contributed positively to the model performance. Alternatively, the MR cutoff value for discarding features (set to 1 in our experiments) could be subjected to hyperparameter tuning. Therefore, future work includes a comparison with other feature selection algorithms.

8.4.2 CHA-Tinnitus

We have 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 and socio-demographics, therefore confirming the multi-factorial nature of tinnitus-related distress. These characteristics were used for the phenotyping in Chapter 5. Additionally, the predictors can be investigated in a followup study of how such characteristics influence treatment success. As expected, predictors 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 also emerged as highly important for the model prediction.

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 analog 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 more fatigued patients who fill some questionnaires less thoroughly.

Our results confirm the intricate interplay between depression and tinnitus-related distress, elucidated by numerous previous studies [66, 81, 109, 168, 231]. For our best model, an ADSL score of more than 20 is associated with an increased predicted risk of tinnitus decompensation (cf. Figure 8.8 (2)), which is close to the cutoff of the clinical relevance of depression [119].

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 a considerable performance improvement. 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)).

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 understand better why all of them were selected and to determine whether some of them could be removed to achieve a better tradeoff 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 various classification algorithm families, both linear and non-linear 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_anxdepression (anxious depressiveness score). While the inclusion of both features contributed to the 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 2701 out of 4117 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 ± 12.2 vs. 51.7 ± 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. [125]. However, psychological treatment approaches are likely to benefit only those who report psychological problems before tinnitus perception or associated with tinnitus perception.

8.4.3 ANEUR

Our classification results are promising, as morphological parameters alone can provide models with moderate power. Because previous studies found that hemodynamic parameters are also predictive [44, 24], future work includes exploring the potential of combining morphologic and hemodynamic features for the classification of rupture status. 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 [61]. We expect that adding these patient characteristics will further improve classification performance.

PDP analysis shows 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} have the highest attribution (see Figure 8.13 (a)). These differ 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 because 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 observe that classification performance is consistently higher for the subset of sidewall aneurysms vs. bifurcation aneurysms and that different parameters are found to have high model attribution. This could be partially due to the relatively small sample size or the already mentioned differences in model families. However, Baharoglu et al. [18] identified significant differences between sidewall and bifurcation aneurysms for morphological parameters and how they 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) is found second most important for ALL - 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 a high classification error in more detail. Here, our goal is to derive descriptions of aneurysms subgroups that are hard to classify to understand reasons for misclassification better and signalize to the medical expert that a manual diagnosis is necessary.

8.5 Conclusion

The previous chapters present methods for subpopulation discovery that yield interpretable results.
In the last years, more accurate black-box models have gained traction. However, additional post-learning steps are required to interpret their behavior and extract actionable insights due to their opacity.

In this chapter, we have complemented the previous chapters' approaches by proposing an end-to-end data analysis workflow for high-dimensional medical data that includes steps for data augmentation, modeling, interleaving model training with feature elimination, and post-hoc analysis of the trained models. The post-mining step of the workflow removes the limitation for medical researchers of being limited to intrinsically interpretable model families by determining the key variables and the corresponding value ranges at the model-, subpopulation-, and observation-level *after* model training. Future work includes investigating the robustness and uncertainty the interpretability methods bring, described as "application-grounded" evaluation by Doshi-Velez and Kim [67].

For some applications, there are already predefined subpopulations, e.g., female and male patients. It is interesting to study their differences regarding the relationship between features and the target variable for a black-box model. We tackle this issue in Chapter 9.

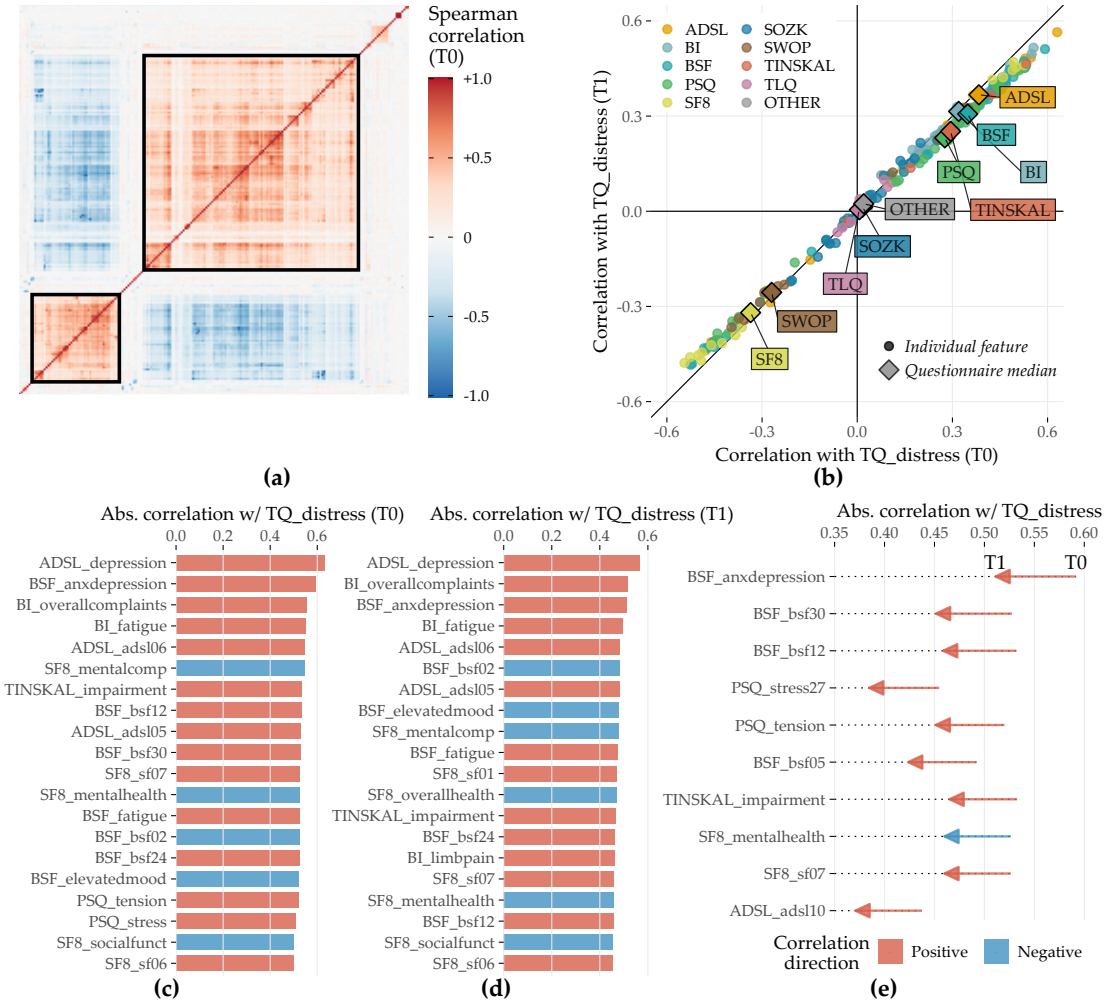


Figure 8.5: **Spearman correlation among predictors and correlation of predictors with TQ_distress in T0 and T1.** (a) The heatmap depicts the correlation coefficients for all pairs of predictors in T0. Predictors are arranged by the result of agglomerative hierarchical clustering with complete linkage. 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 T1 (y-axis). The diamond symbol represents the median correlation of the predictors from the same questionnaire. (c) Top 20 predictors that exhibit the highest absolute correlation with TQ_distress in T0. (d) Top 20 predictors which exhibit the 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.

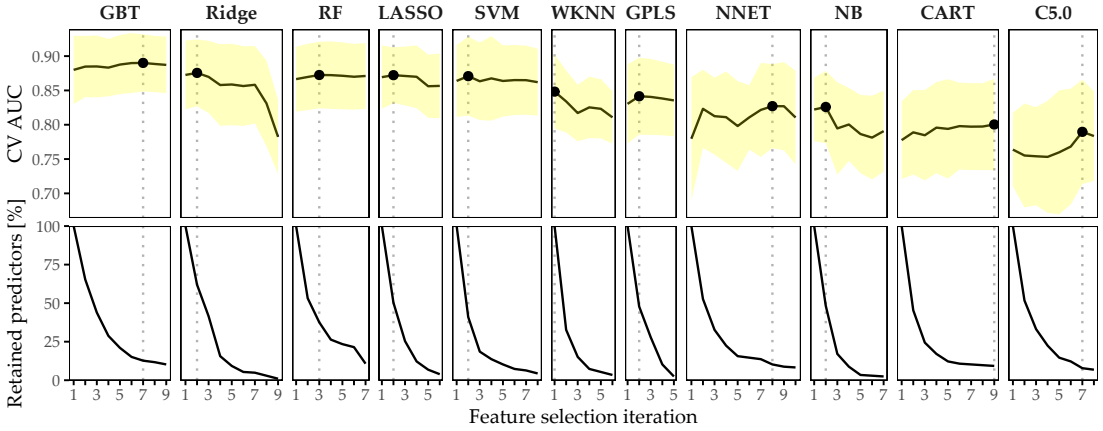


Figure 8.6: **Classification results for CHA-Tinnitus**. Average cross-validation AUC and relative number of retained predictors for each classifier with optimal hyperparameter configuration and each feature selection iteration. Yellow ribbons depict standard deviation. Points highlight each classifier's run with maximum AUC. Classifiers are ordered by their maximum AUC from left to right.

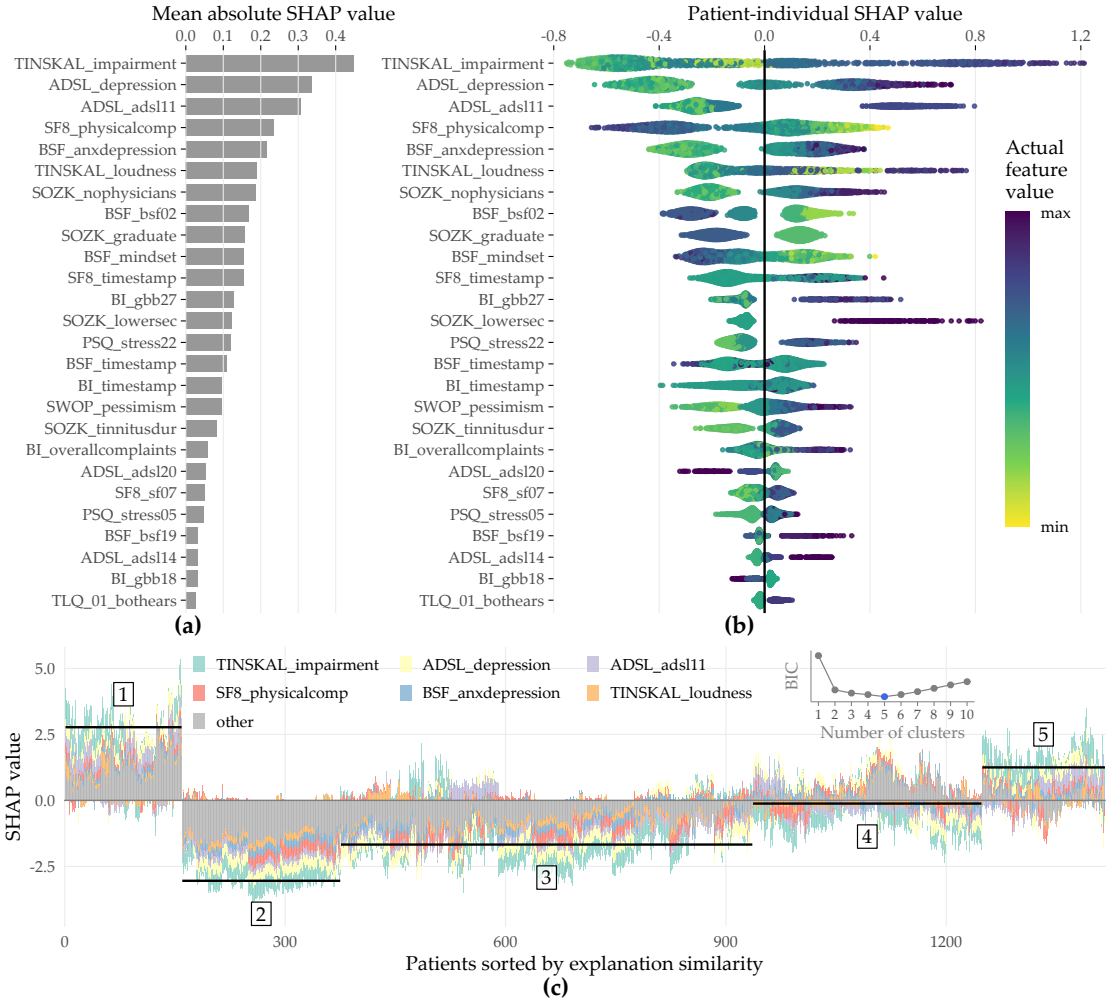


Figure 8.7: **SHAP analysis results for the best model (GBT, feature elimination iteration i = 7)**. (a) Global feature importance based on the mean absolute SHAP magnitude over all observations. Values depict the absolute change in log odds where higher values indicate higher feature attribution towards the model. (b) Patient-individual SHAP values. A point 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 high-density regions (similar to a violin plot), i.e., there are more patients with similar SHAP values. Actual predictor values are mapped to point color. (c) Stacked patient-individual SHAP values for the six predictors with the 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 the Bayesian information criterion (BIC) is minimal for this number of clusters.

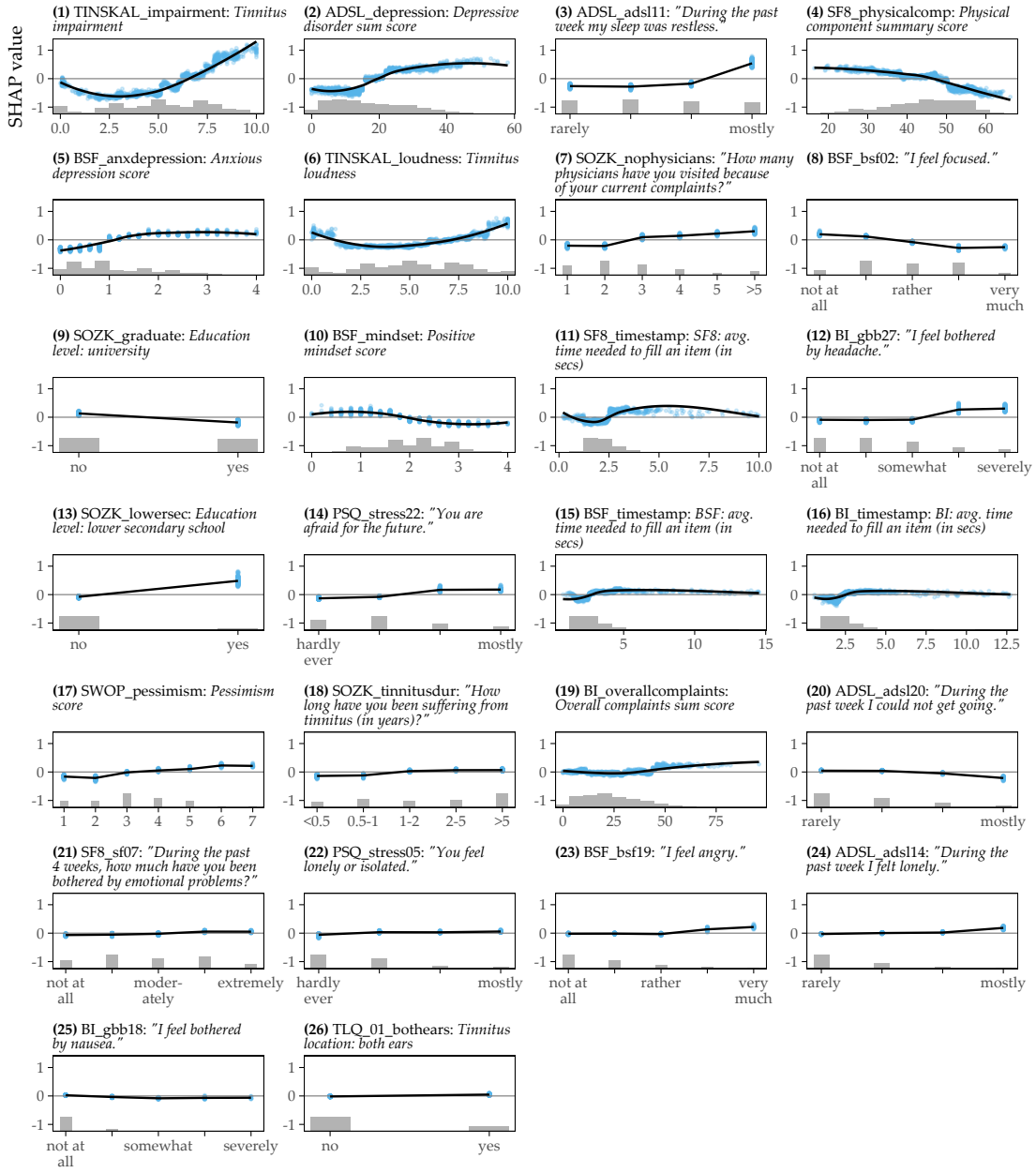


Figure 8.8: **SHAP feature dependence.** The relationship between the actual values of a predictor (x-axis) and corresponding SHAP values (y-axis) is shown as points representing a patient and as locally weighted scatterplot smoothing (LOWESS) [52] curves indicating the overall trend. Predictors are ordered by mean absolute SHAP value (see Figure 8.7 (a)). Gray histograms and bar charts depict the distributions of the predictors.

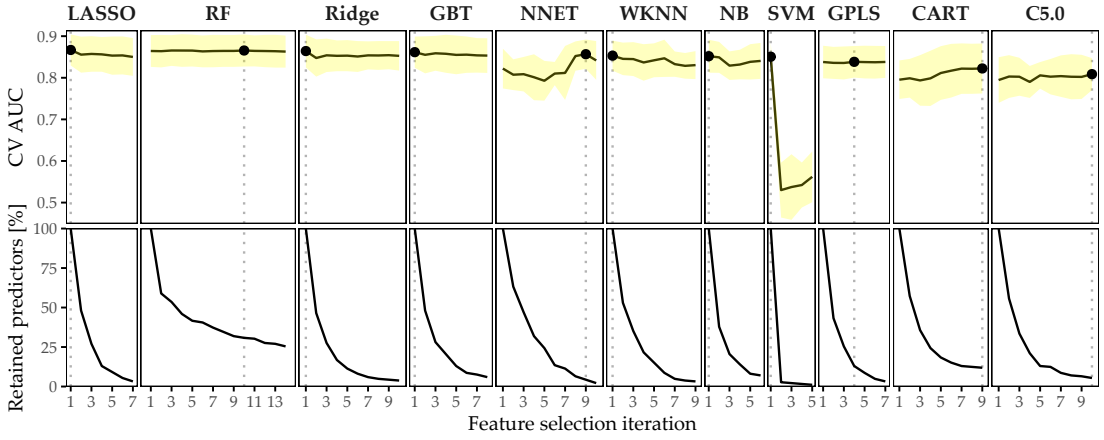


Figure 8.9: **Classification results for CHA-Depression.** Average cross-validation AUC and relative number of retained predictors for each classifier with optimal hyperparameter configuration and each feature selection iteration. Yellow ribbons depict standard deviation. Points highlight each classifier's run with maximum AUC. Classifiers are ordered by their maximum AUC from left to right.

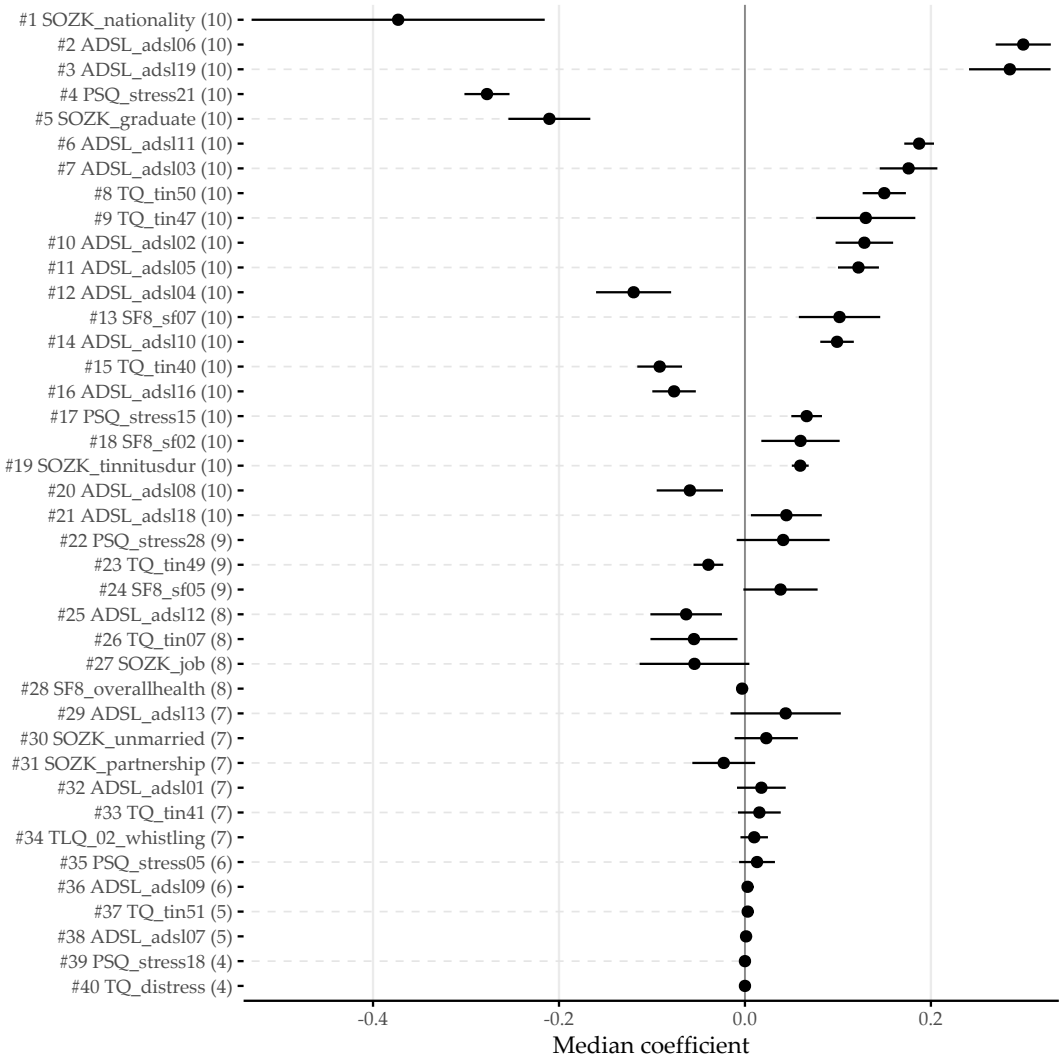


Figure 8.10: **Coefficients of LASSO model.** Cross-validation (CV) median ± median absolute deviation (line ranges) of coefficients for the best LASSO model (i = 1). The frequency of non-zero coefficients in 10-fold CV is given in parentheses right to the predictor name. From 185 features in total, 40 features exhibit a non-zero model coefficient for at least one CV fold.

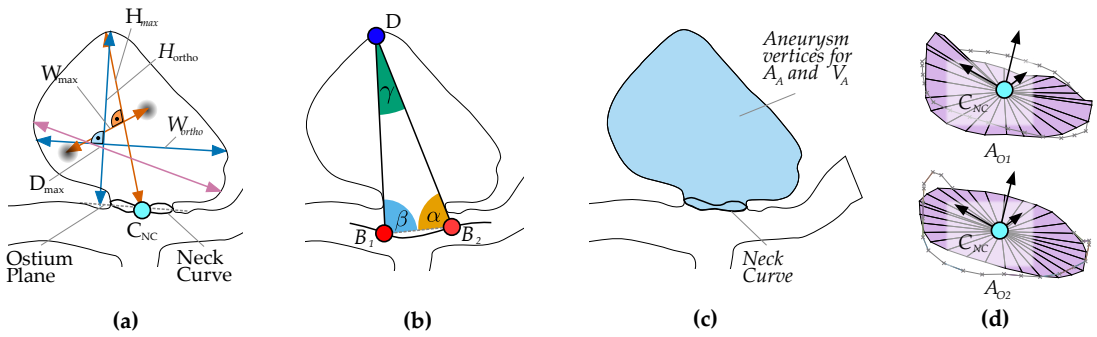


Figure 8.11: **Illustration of the extracted morphological features.** (a) Features that describe aneurysm width, height, and diameter. (b) The angles α, β and γ are extracted from the base points B_1, B_2 and the dome point D. (c) After separating the aneurysm from its parent vessel via the neck curve, 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}.

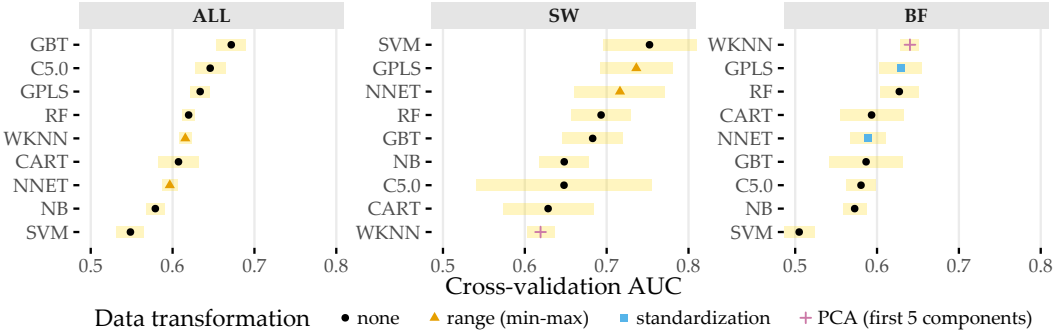


Figure 8.12: **Classification results for ANEUR.** For each combination of data subset and classification algorithm, the performance of the run with the preprocessing transformation that achieves the highest AUC is shown. SW = sidewall; BF = bifurcation.

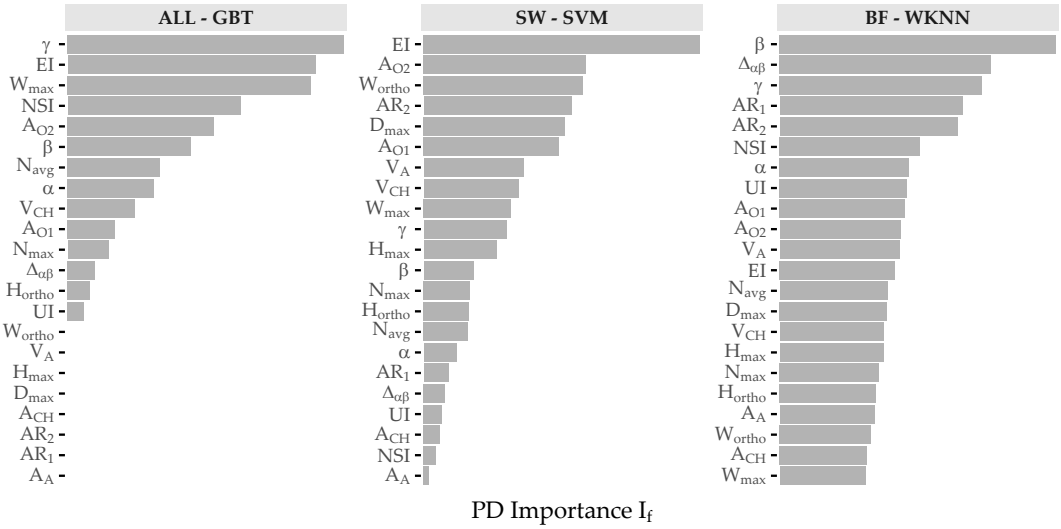


Figure 8.13: **Relative PD importance (ANEUR).** PD importance for the best model of each data subset. Values are relative to the maximum PD importance. SW = sidewall; BF = bifurcation.

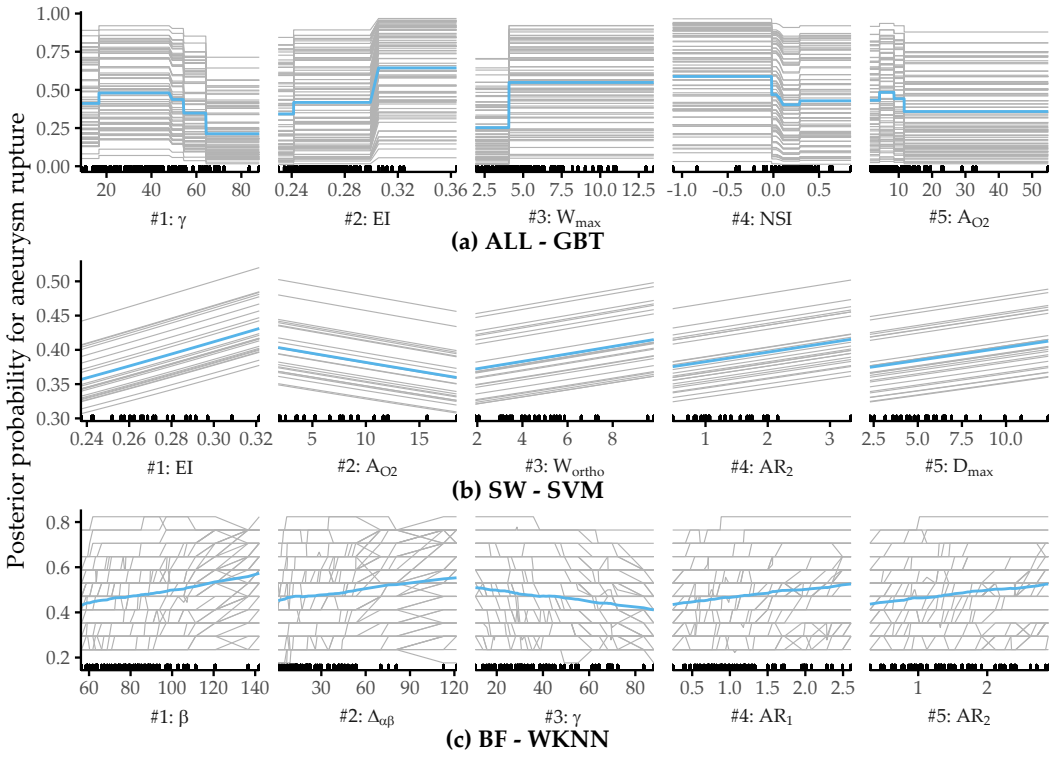


Figure 8.14: **Relative PD importance (ANEUR).** PD importance for the best model of each data subset. Values are relative to the maximum PD importance. SW = sidewall; BF = bifurcation.

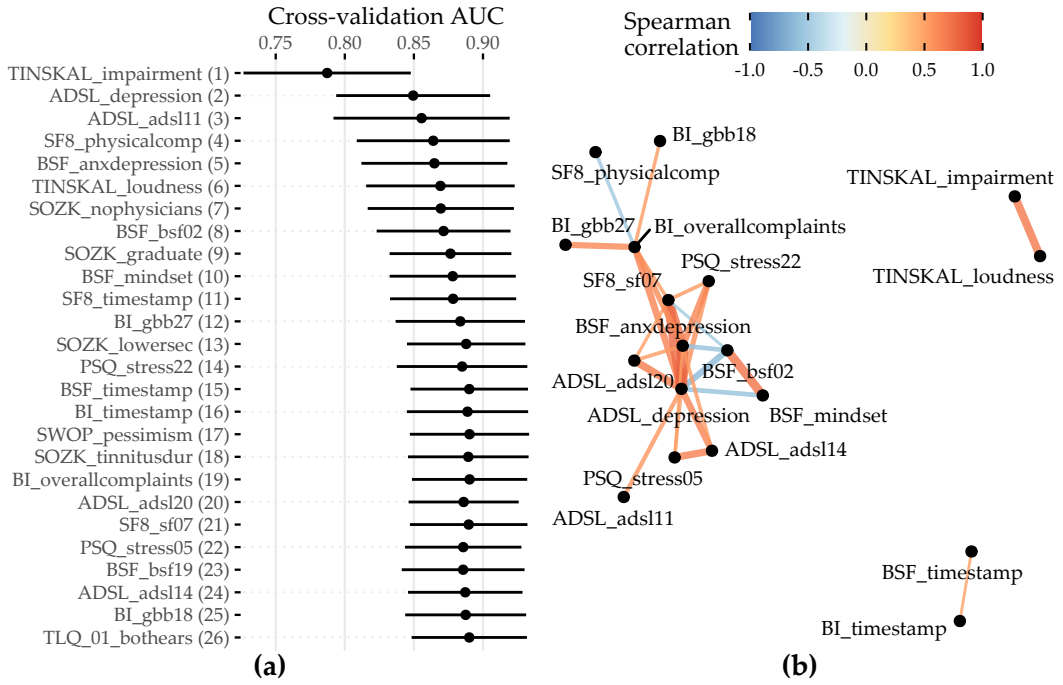


Figure 8.15: **Cumulative feature contribution and correlation network.** (a) Cross-validation AUC (average ± standard deviation) of a GBT model trained on the feature subset comprising the predictors denoted on the y-axis up to that iteration. The ordering of features is according to the mean absolute SHAP value (cf. Figure 8.7 (a)). (b) Network illustrating 3 groups of features among the 26 selected predictors of the best model with high intra-group correlation ($|\rho| \geq 0.5$). Eight predictors (predominantly from SOZK) without any moderate to high pairwise correlation are not shown.

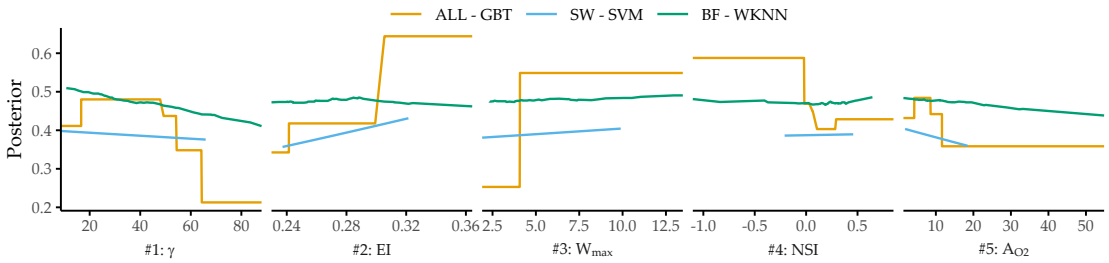


Figure 8.16: **PDP curves for top 5 predictors on ALL - GBT for each data subset's best model.**

Chapter 9

Subpopulation-Specific Learning and Post-Hoc Model Interpretation

Brief Chapter Summary

We present a workflow to examine differences between *two* a priori defined subpopulations in temporal data. For this purpose, we derive a post-hoc interpretation measure to visually assess the difference in the features' relationship with the predicted target variable between two subpopulations. We validate our approach on two data samples and the target variables tinnitus distress (CHA), depression (CHA), and liver fat concentration (SHIP). We determine gender-specific differences, i.e., we separate between discriminating features (i) for both, (ii) for one of the two, and (iii) neither of the subpopulations.

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), pp. 1-11. DOI: 10.3389/fnins.2020.00487.

The previous chapter describes a workflow for the post-hoc analysis of models built on the entire population. For several medical applications, there are either already established subpopulations that need further investigation, or it is unclear whether there are indeed differences between predefined subgroups of interest.

This chapter presents a workflow to explore how two disjoint subpopulations differ concerning their most predictive features. First, we build machine learning models that separate between two subpopulations to identify informative features associated with either subpopulation. Then, for each subpopulation separately, we train models that predict the value of a target variable. We use a quantitative and a qualitative mechanism to compare differences in feature importance between the subpopulations. We validate our workflow on CHA and SHIP data samples. For

CHA, we investigate questionnaire items and scores predictive of tinnitus-related distress and depression. For SHIP, we identify variables from the baseline examinations that are potentially long-term determinants of fatty liver measured at the second follow-up ten years later.

This chapter is organized as follows. In Section 9.1, we motivate for subpopulation-specific model interpretation by discussing gender differences in tinnitus. Section 9.2 presents the workflow. Section 9.3 describes the measure to quantify and visualize subpopulation-specific model differences. Section 9.4 explains the validation setup, including learning tasks, selected learning algorithms, and dataset-specific preprocessing steps. In Section 9.5, we report our results and main findings. The chapter closes with a summary and a brief discussion in Section 9.6.

9.1 Motivation and Comparison to Related Work

One example where the relationship between subpopulation membership on the target variable is not well understood yet is gender differences in tinnitus patients. Female gender has been identified as an important risk factor for psychological comorbidities in many studies: women show higher prevalence rates regarding depression, anxiety, and other psychosomatic diseases [205, 213, 189, 147, 192, 10, 169]. However, previous studies presented conflicting results on the relationship between gender and tinnitus severity and distress. Some studies found no gender differences with respect to tinnitus severity [71], annoyance [215], and similar tinnitus-related scores [193]. Others found higher tinnitus distress in women [244], higher loudness and annoyance in men [129], a high association of severe tinnitus with suicide attempts only in women [182], and a high correlation of tinnitus severity with life quality, depression, and stress only in men [114]. Overall, there is no consensus on gender-specific determinants of prevalence rates or accompanying symptoms of chronic tinnitus such as depression or anxiety. However, gender differences in psychological response profiles and coping strategies could substantially influence tinnitus chronification and treatment success rates [273]. Understanding gender differences may therefore facilitate a more detailed identification of symptom profiles, increase treatment response rates, and help provide access for vulnerable populations who may be less visible in the clinical setting.

9.2 Workflow

We present a workflow consisting of a modeling component and a post-modeling component, extending the approach presented in Chapter 8. More specifically, we (i) build machine learning models capable of predicting the values of a target variable while performing hyperparameter optimization in a cross-validation scheme, and (ii) use post-hoc interpretation mechanisms to identify variables that contributed most to the prediction of the best model. We deviate from the workflow presented earlier because we are interested in comparing a priori defined disjoint subpopulations instead of determining them.

We proceed as follows. First, we train a model for each population where subpopulation membership is the target variable. Then, we build models for each subpopulation separately and compare model reliance values between the best performing models. More specifically, we compare which variables appear to be important (i) for both subpopulations, (ii) for one of the two subpopulations, or (iii) for neither subpopulation. For example, we split a dataset into two non-overlapping based on the variable defining the subpopulations and validate our workflow by examining factors that discriminate between subpopulations and factors predictive for the response variable in both or either of the two subpopulations.

9.3 Comparing Differences in Feature Importance between Two Subpopulations

To measure feature-individual attributions to a model's predictions and to compare them between two subpopulations, we use model reliance (MR; [79]), which is described in the context of iterative feature elimination in Section 8.2.3. Recall that $MR(f, \zeta)$ is a permutation-based variable importance measure that calculates the increase in the error of a model ζ when the values of the variable of interest f are randomly shuffled within the training set. If f is important for the prediction of ζ, $MR(f, \zeta) > 1$. If random permutation of the feature values leads to a higher performance of the model, then the feature's attribute to model quality is low, whereupon $MR < 1$.

For models that predict subpopulation membership, we rank features by MR value and report on the most important features. For subpopulation-specific models, we use scatterplots (see Figure 9.1) depicting a feature's contribution (as MR value) to the model Model_1 (x-axis), which is trained on one subpopulation, and to the model Model_2 (y-axis), trained on the other subpopulation. Features that contribute equally to both models are on the diagonal line, while higher MR values for one model are further from the diagonal. The variables with the highest average MR score or the highest difference magnitude between subpopulation-specific MR scores are colored and labeled.

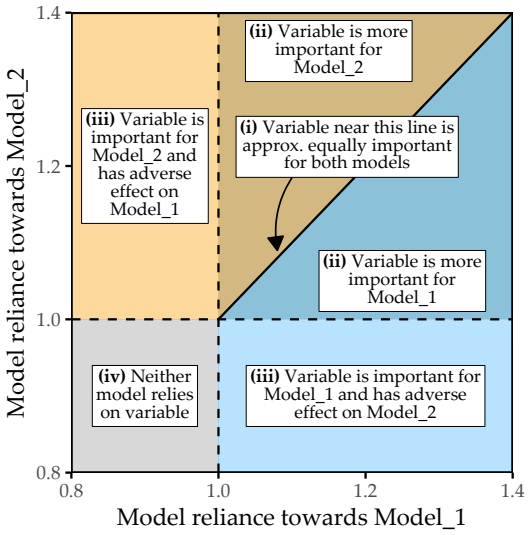


Figure 9.1: **Subpopulation-specific variable importance.** The position of a point represents the model reliance score of a variable for the best model trained on the respective subpopulation, denoted as Model_1 (x-axis) and Model_2 (y-axis). Higher values represent a higher attribution of a variable relative to the model prediction. There are four characteristic areas: (i) important variables with similar attributions to Model_1 and Model_2; (ii) important variables with higher attribution to one of the subpopulation-specific models; (iii) variables important to *either* Model_1 *or* Model_2 but adversarial to the other model; (iv) variables that are adversarial to both models.

9.4 Learning Tasks and Evaluation Setup

We validate our workflow on the Charité tinnitus patient dataset (CHA, Section 2.2.2) and the SHIP dataset (Section 2.2.1). We define five *learning tasks* (LT) for the following response variables and subpopulations:

- LT 1: gender (CHA)
- LT 2: tinnitus-related distress (CHA)
  - LT 2a: female subpopulation
  - LT 2b: male subpopulation
- LT 3: depression (CHA)
  - LT 3a: female subpopulation
  - LT 3b: male subpopulation
- LT 4: gender (SHIP)
- LT 5: liver fat concentration at second follow-up (SHIP)
  - LT 5a: female subpopulation
  - LT 5b: male subpopulation

LT 1 and LT 4 have gender as the target variable, and characteristics are identified that are predictive for one of the genders. For each learning task, we use variables from the first study as predictors. For the learning tasks LT 2, LT 3, and LT 5, we build separate models for each of the two gender subpopulations. We refer to these models as "F_model" and "M_model", respectively; the learning task is explicitly stated if it cannot be inferred from the context.

Selection of algorithms and evaluation. Based on their encouraging performances relative to other classifiers in Chapter 8, we use the following five algorithms: least absolute shrinkage and selection operator (LASSO [84]), RIDGE [134], support vector machine (SVM [31]), random forest (RF [35]) and gradient boosted trees (GBT [85]), see Section 8.2.2 for a description. We use 10-fold cross-validation to evaluate model generalization performance and perform a grid search for hyperparameter selection (cf. listing of parameter candidates in Table 8.1). We choose the evaluation measures based on the type of the target variable. For LT 1, we employ accuracy and sensitivity for each gender. For LT 2, we use root mean squared error (RMSE) and the coefficient of determination R^2, defined as $R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}$, where $\bar{y}$ is the average response value. Higher values are better for all measures except RMSE.

For each learning task, we define a baseline performance. For the classification problem of LT 1 and LT 4, the baseline is equal to a model that always predicts the majority class over all training observations. Similarly, for the regression problems of LT 2, 3, and 5, the average value of a target variable over all training observations is used to predict every test instance's response value.

Data preparation for CHA. For CHA, we use baseline data collected before the start of therapy. To make the learning tasks nontrivial, we remove variables from the same questionnaire as the response variable for each task. For example, for LT 2, we exclude all variables from the TQ questionnaire because the target variable is calculated from them. We include 1628 patients (828 female, 800 male) with complete data for the ADSL, PSQ, SF8, SOZK, TINSKAL, TQ, and TLQ questionnaires (cf. Table 2.1). The selection of these questionnaires is motivated to

obtain a comprehensive assessment of tinnitus, comorbid conditions (e.g., depression), general quality of life, and socio-demographic data. For CHA and SHIP, multinomial variables, i.e., variables that take three or more symbolic values, such as reported gender, marital status, and education level, are encoded as dummy variables. For example, smoking status smoking_s0 can be one of the following values: 0 = never a smoker, 1 = ex-smoker, 2 = current smoker. The dummy variable smoking_s0_1 then indicates whether a study participant is an ex-smoker. To avoid multicollinearity, we remove the first dummy of each variable, leaving only n-1 dummies. For the smoking status example, we keep smoking_s0_1 and smoking_s0_2, and we remove smoking_s0_0. Finally, a total of 181 variables from the baseline measurements are used as predictors, including responses to individual questionnaire items, subscale scores, total scores, and, for each questionnaire, the average time taken to complete an item. Tinnitus-related distress is measured by the TQ total score (TQ_distress). The severity of depression is measured by the ADSL total score (ADSL_depression).

Data preparation for SHIP. For the SHIP data, we consider only the variables recorded in SHIP-0 and use the liver fat concentration (liverfat_s2) measured via MRT in SHIP-2. We use the same subset of 886 labeled participants as in Chapter 4, of which 460 are female and 426 male. We remove the variables related to the ultrasound diagnosis of hepatic steatosis, stea_s0 and stea_alt75_s0, because we have already identified their high correlation to the target in Chapters 3 and 4. Furthermore, we remove "near-zero" variance variables where the most frequent value occurs at least 19 times more frequently than the second most frequent value. Typical examples include the ATC_* medication variables where few participants report taking them. Variables with a variance near zero can lead to resampling problems because some of the resamples may have constant values for that variable. Besides, it is difficult to infer significant correlations from them, as it is unclear whether the measured effects can be generalized to the overall population or whether they are just an artifact of this small and thus unrepresentative sample.

Finally, we remove highly correlated features using the algorithm of Kuhn and Johnson [162]. We first compute the Pearson correlation coefficient for each pair of features. For feature pairs with an absolute correlation value of $r \geq 0.9$, we keep the feature that has the lower average correlation with the other features. We repeat this process until none of the correlation values exceed the specified threshold. Of the original 350 variables, 118 remain to be used for modeling. Since determining the appropriate type of missingness for each variable is beyond our workflow's scope, we assume that missingness occurs completely at random (MCAR). Missing values are thus imputed by random sampling with replacement.

9.5 Validation on Two Datasets

9.5.1 Results for CHA

Distribution of the target variables. Figure 9.4 shows the target variables' distributions for the CHA learning tasks (LTs) 1-3. There are slightly more female than male patients. In general, female patients report higher levels of tinnitus-related distress (median ± median absolute deviation (MAD) 39.0 ± 17.8 vs. 35.5 ± 20.0) and depression (18.0 ± 11.9 vs. 14.00 ± 11.9). TQ_distress and ADSL_depression are right-skewed in each subpopulation. Using the TQ cutoff of 46 for tinnitus distress [98], 34.1% of females and 30.8% of males show decompensated tinnitus. Using the ADSL cutoff of 15 for depression severity [119], 57.4% of female and 45.0% of male subjects exhibit clinical depression.

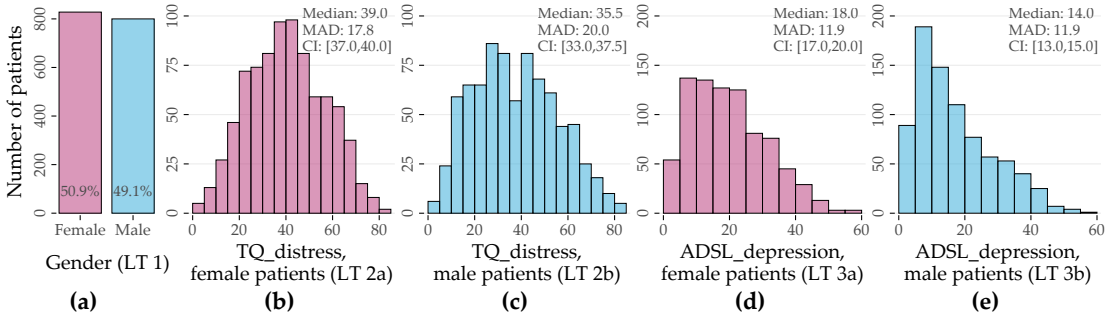


Figure 9.2: **Distribution of target variables (LT 1-3)**. For the numerical targets, median, median absolute deviation (MAD), and non-parametric 95% confidence interval (CI) using bootstrap sampling [64] with 2000 samples are presented.

Table 9.1: **Classifier performance (LT 1).** Performance values are cross-validation mean ± standard deviation. The best performance for each measure is highlighted in bold. Sens. = Sensitivity.

Algorithm	**Accuracy (%)**	**Sens. female (%)**	**Sens. male (%)**
LASSO	71.3 ± 3.0	69.5 ± 4.2	**73.2 ± 4.9**
Ridge	**72.2 ± 2.9**	**71.4 ± 5.5**	73.0 ± 4.3
SVM	71.8 ± 3.7	70.9 ± 5.2	72.7 ± 6.0
RF	68.6 ± 5.2	68.7 ± 6.0	68.6 ± 7.1
GBT	70.4 ± 3.4	70.9 ± 5.1	69.7 ± 6.2
Baseline	*50.9 ± 2.9*	*100.0 ± 0.0*	*0.0 ± 0.0*

Tables 9.1-9.5 provide an overview of the generalization performances of each method for each learning task.

Learning task 1 (CHA, gender classification). Ridge achieves best cross-validation accuracy (mean: 72.2% ± standard deviation: 2.9%) with a sensitivity of 71.4% ± 5.5% for female patients and 73.0 ± 4.3% for male patients. Figure 9.3 illustrates the item response frequencies for the variables among the top 5% with respect to model reliance (MR), i.e., the 8 variables with the highest attribution to the model prediction. For each variable, the horizontal legend shows the corresponding text of the questionnaire item. The vertical axis shows the responses to that item, and the horizontal axis depicts the relative frequency by gender. These frequencies are shown as bars, red-violet for female patients and blue for male patients. A difference in the length of the two bars for the same answer means that the percentage of giving that response is different for each gender; thus, the variable is contributing to class separation.

The item ADSL_adsl17 (Figure 9.3 (a)) is the most discriminating variable for the model (MR = 1.167): while 16% of female patients report having had crying spells either "mostly" or "occasionally" in the past week, only 4% of male patients do; they predominantly give the answer "rarely" (86.2%). Female patients tend to express higher levels of worry (see Figures 9.3 (b) and 9.3 (f)) and subjective stress (see Figure 9.3 (h)). Besides, differences in tinnitus quality were found: More than half (52.4%) of all male patients report the tinnitus sound (MR = 1.056) as "whistling", which is substantially more frequent than in female patients (35.6%), who describe their tinnitus as "rustling" noise more often (33.3%) than male patients (22.9%).

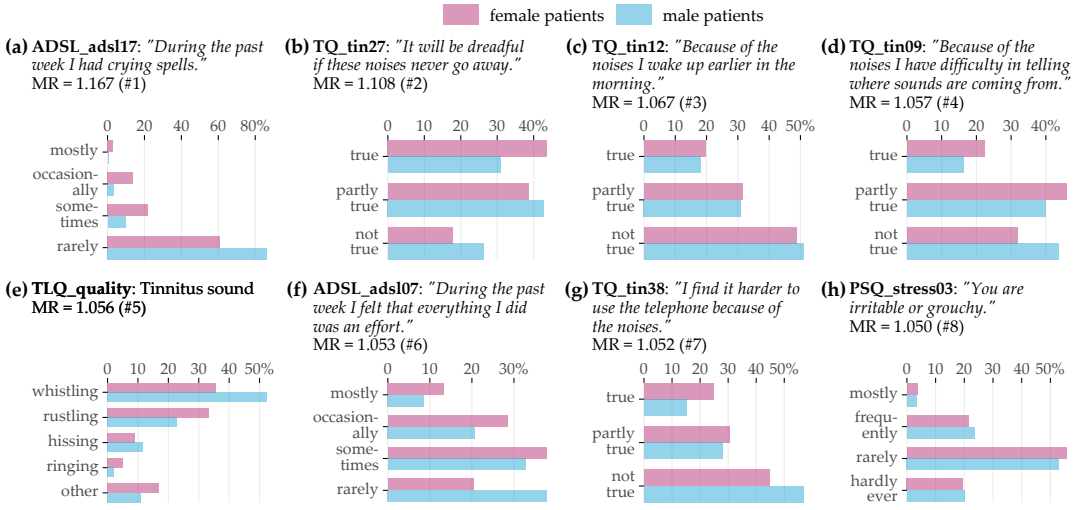
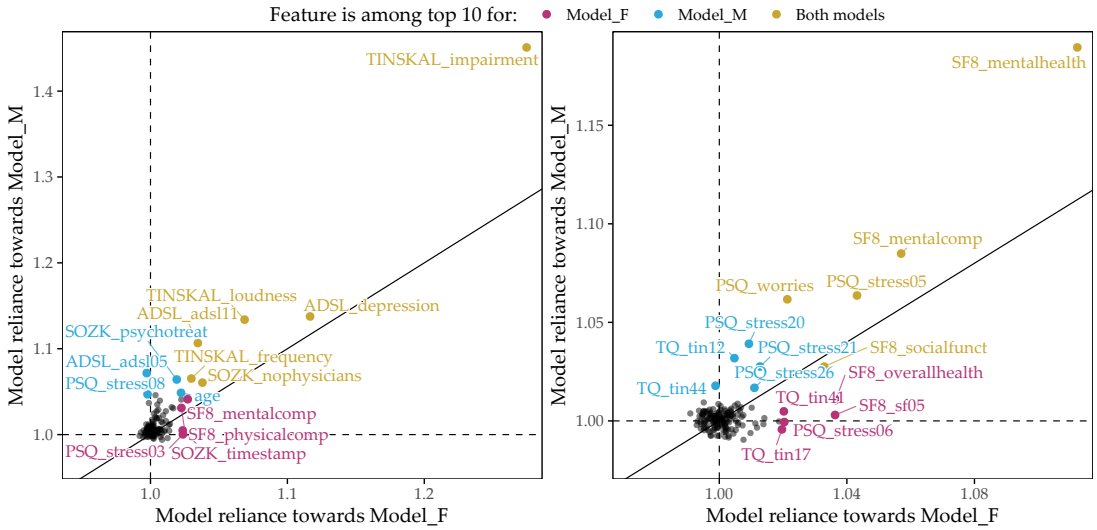


Figure 9.3: **Top 8 variables on gender (LT 1).** Gender-specific item response frequencies for the top 5% variables with the highest attribution towards model prediction according to model reliance (MR).

Table 9.2: **Regression model performance (LT 2a, LT 2b).** Performance values are cross-validation mean ± standard deviation. The best performance for each measure is highlighted in bold.

	LT 2a (female)		LT 2b (male)	
Algorithm	**RMSE**	**R^2**	**RMSE**	**R^2**
LASSO	11.55 ± 0.70	0.50 ± 0.04	10.59 ± 0.98	0.65 ± 0.05
Ridge	11.59 ± 0.63	0.50 ± 0.04	10.72 ± 1.05	0.64 ± 0.06
SVM	11.97 ± 0.51	0.46 ± 0.03	11.21 ± 1.02	0.61 ± 0.06
RF	11.38 ± 0.74	0.51 ± 0.05	10.58 ± 1.01	0.65 ± 0.06
GBT	**10.92 ± 0.68**	**0.55 ± 0.04**	**10.11 ± 1.12**	**0.68 ± 0.06**
Baseline	*16.22 ± 1.38*	*0.00 ± 0.00*	*17.77 ± 1.00*	*0.00 ± 0.00*

Learning task 2 (CHA, tinnitus distress prediction). For LT 2, we ran the five algorithms once for the female patients (LT 2a) and once for the male patients (LT 2b). Table 9.2 shows RMSE and R^2 for each algorithm. For both LT 2a and LT 2b, GBT exhibits the best performance in terms of RMSE (LT 2a: 10.92 ± 0.68, LT 2b: 10.11 ± 1.12) and R^2 (LT 2a: 0.55 ± 0.04, LT 2b: 0.68 ± 0.06). It is noticeable that GBT and the other models are slightly more accurate for male patients than for females. The highest MR feature attribution is achieved by TINSKAL_impairment, i.e., the TINSKAL visual analog scale for tinnitus impairment. Figure 9.4 (a) illustrates that MR scores are higher for male patients (MR = 1.42 vs. 1.24). Furthermore, the variables ADSL_depression (depression), ADSL_adsl11 (sleep problems), and TINSKAL_loudness (tinnitus loudness) were found to be important for both models. It is noteworthy that the MR scores of most of the 120 variables are close to 1, which is visualized by the clump of points in Figure 9.4 (a). In fact, only 8 variables exhibit a substantial attribution with MR $>$ 1.05 for either of the gender-specific models.



(a) Learning task 2: TQ_distress **(b)** Learning task 3: ADSL_depression

Figure 9.4: **Juxtaposition of variable importance (LT 2 and LT 3)**. Each scatterplot shows the model reliance (MR) score for each predictor for the best model for each of the subpopulations of female (x-axis) and male (y-axis) patients and for each learning task; see Figure 9.1. Variables among the top 10 highest-ranking variables by MR in the F_model (red-violet), M_model (blue), or both models (yellow) are highlighted. (a) LT 2 (CHA, response: tinnitus-related distress); (b) LT 3 (CHA, response: depression severity).

Learning task 3 (CHA, depression prediction). For depression severity, LASSO provides the best model for both female patients (RMSE = 5.80 ± 0.73; R^2 = 0.74 ± 0.06) and male patients (RMSE = 5.10 ± 0.38; R^2 = 0.81 ± 0.03), as depicted in Table 9.3. Similar to LT2, the RMSE and R^2 estimates for the models are consistently better for the subgroup of male patients. Figure 9.4 (b) shows that the mental health indicator SF8_mentalhealth is the most important predictor for both the F_model and the M_model. Furthermore, features measuring subjective stress (PSQ_stress05: *"You feel lonely or isolated."*), worry (PSQ_worries score), and vitality (SF8_sf05: *"How much energy have you had in the last 4 weeks?"*) contribute substantially to the predictions of both genders' models.

9.5.2 Results for SHIP

Learning task 4 (SHIP, gender classification). Table 9.4 shows that LASSO performs best in terms of accuracy (99.4% ± 0.8%) and sensitivity for the male class (99.5% ± 1.0%), while GBT achieves the highest sensitivity for the female class (99.6% ± 0.9%). All regression models outperform the baseline, which constantly predicts the female majority class (51.9% ± 4.8%). Figure 9.5 depicts the distributions of the 4 features with an MR score of at least 1.05 for the LASSO model. The two anthropometric measures, waist circumference (som_tail_s0) and hip circumference (som_huef_s0), appear to be predictive of the gender of the study participant. From Figure 9.5 (a), it can be inferred that, in general, men have a higher waist circumference than women (median: 92.2 cm vs. 77.5 cm). While the median hip circumferences are similar (f: 99.8 cm, m: 100.0 cm), it can be seen in Figure 9.5 (d) that the distribution of women has a wider spread and more values beyond the distribution tails of men, i.e., values below 86 cm and above 123 cm. The second most important feature, gfr_mdrd_s0, is the glomerular filtration rate, a measure of overall renal function that describes the flow rate of filtered fluid through the

Table 9.3: **Regression model performance (LT 3a, LT 3b).** Performance values are cross-validation mean ± standard deviation. The best performance for each measure is highlighted in bold.

	LT 3a (female)		LT 3b (male)	
Algorithm	**RMSE**	**R^2**	**RMSE**	**R^2**
LASSO	**5.80 ± 0.73**	**0.74 ± 0.06**	**5.10 ± 0.38**	**0.81 ± 0.03**
Ridge	5.88 ± 0.65	**0.74 ± 0.06**	5.14 ± 0.38	0.80 ± 0.03
SVM	6.12 ± 0.72	0.71 ± 0.06	5.29 ± 0.38	0.79 ± 0.03
RF	6.02 ± 0.62	0.72 ± 0.05	5.29 ± 0.46	0.79 ± 0.04
GBT	6.10 ± 0.65	0.72 ± 0.06	5.16 ± 0.42	0.80 ± 0.04
Baseline	*11.36 ± 0.78*	*0.00 ± 0.00*	*11.52 ± 0.47*	*0.00 ± 0.00*

Table 9.4: **Classifier performance (LT 4).** Performance values are cross-validation mean ± standard deviation. The best performance for each measure is highlighted in bold. Sens. = Sensitivity.

Algorithm	**Accuracy (%)**	**Sens. female (%)**	**Sens. male (%)**
LASSO	**99.4 ± 0.8**	99.4 ± 1.0	**99.5 ± 1.0**
Ridge	98.2 ± 1.9	98.1 ± 2.0	98.3 ± 2.3
SVM	99.1 ± 0.9	98.9 ± 1.1	99.3 ± 1.2
RF	96.6 ± 2.5	98.3 ± 1.9	94.9 ± 4.1
GBT	99.3 ± 0.8	**99.6 ± 0.9**	99.1 ± 1.2
Baseline	*51.9 ± 4.8*	*100.0 ± 0.0*	*0.0 ± 0.0*

kidney. Figure 9.5 (b) shows that gfr_mdrd_s0 is generally higher in men, consistent with the literature [115]. The third most important feature, crea_s_s0 (Figure 9.5 (c)), measures serum creatinine concentration, another indicator of renal function, which is higher in men [78].

Learning task 5 (SHIP, prediction of liver fat concentration). The target variable liver fat_s2 is highly right-skewed in each gender subpopulation (Figure 9.6 (a)). For example, whereas the lower half of the females' distribution lies between 1.18% and 3.29%, the upper half has a much wider spread, ranging between 3.29% and 41.8%. Males have a higher median liver fat concentration (5.57%) than females (3.29%). In the female subpopulation (LT 5a), LASSO performs best in terms of RMSE (5.76 ± 1.05) and R^2 (0.23 ± 0.12), whereas GBT achieves a minimum RMSE (6.02 ± 1.11) and a maximum R^2 (0.20 ± 0.16) in the male subpopulation (LT 5b), see Table 9.5. All regression models for LT 5a outperform the baseline predicting subpopulation mean liver fat concentration. Only SVM performs worse on the RMSE than the baseline for LT 5b.

Figure 9.6 (b) visualizes the LASSO models' MR scores for the female and male subpopulations. Waist circumference (som_tail_s0) and serum uric acid concentration (hrs_s_s0) are the only predictors that appear among the top 10 features in both subpopulations. Of note, several features are considerably responsible for model predictions in one subpopulation but not the other, as illustrated by the points near the horizontal and vertical dashed lines. Recall that these lines indicate MR = 1, which expresses that a feature neither contributes to nor hinders model performance.

Waist circumference (som_tail_s0) and serum uric acid concentration (hrs_s_s0) are the only

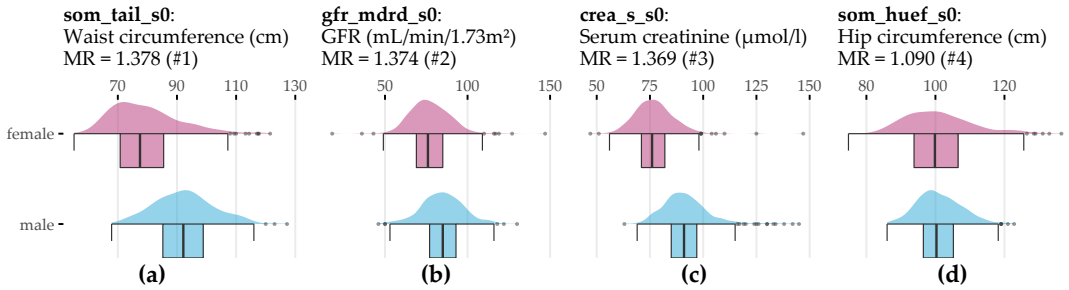


Figure 9.5: **Top 4 variables on gender (LT 4).** Gender-specific item response frequencies for the 4 variables with the highest attribution towards model prediction according to model reliance (MR). For crea_s_s0, an outlier (creas_s_s0 = 281) was removed to preserve plot readability. GFR = glomerular filtration rate.

Table 9.5: **Regression model performance (LT 5a, LT 5b).** Performance values are cross-validation mean ± standard deviation. The best performance for each measure is highlighted in bold.

	LT 5a (female)		**LT 5b (male)**	
Algorithm	**RMSE**	**R^2**	**RMSE**	**R^2**
LASSO	**5.76 ± 1.05**	**0.23 ± 0.12**	6.18 ± 1.18	0.16 ± 0.22
Ridge	5.89 ± 1.09	0.20 ± 0.11	6.39 ± 1.13	0.11 ± 0.17
SVM	6.03 ± 1.09	0.17 ± 0.13	6.76 ± 1.29	0.02 ± 0.09
RF	5.88 ± 1.06	0.20 ± 0.10	6.18 ± 1.06	0.16 ± 0.17
GBT	6.02 ± 0.88	0.14 ± 0.16	**6.02 ± 1.11**	**0.20 ± 0.16**
Baseline	*6.48 ± 1.04*	*0.00 ± 0.00*	*6.71 ± 1.21*	*0.00 ± 0.00*

predictors that appear among the top 10 features in both subpopulations. The feature metabolic syndrome (metsyn_s0_1) appears to be predictive only for the female subpopulation. Indeed, the Pearson point biserial correlation of metsyn_s0_1 and liverfat_s2 is $r = 0.44$ for the female subpopulation, which is considerably larger than for the male subpopulation, where $r = 0.22$. Similarly, the feature alat_s_s0, which measures alanine aminotransferase (ALAT) concentration, is more informative of liver fat concentration in the male subpopulation ($r = 0.38$ in men vs $r = 0.16$ in women). Furthermore, important characteristics for women are increased waist circumference categorization (waiidf_s0_1; waist circumference ≥ 80 cm; cutoff is 94 cm for men), serum triglyceride concentration (tq_s_s0), heart rate (heartr_s0), sleep problems (sleepp_s0_1), waist-to-hip ratio (whratc_s0_1), antihypertensive medication (antihyp_s0_1), and ferritin concentration (ferri_s0). In men, other features with the highest model attribution include restless legs syndrome (rlegs_s0_1), serum creatinine concentration (crea_s0_s0), body mass index (som_bmi_s0), diastolic blood pressure (diabp_s0), smoking status (smoking_s0_1), the genetic marker rs11597086 (gx_rs11597086_2) which is associated with ALAT concentration [300], and the "SF-12 Physical and Mental Health Summary Scale" score (mcs_sf12_s0; [38]). In females (males), for 10 (16) of 116 features, it holds that MR ≥ 1. These numbers are identical to the number of features with a nonzero coefficient in the respective LASSO models.

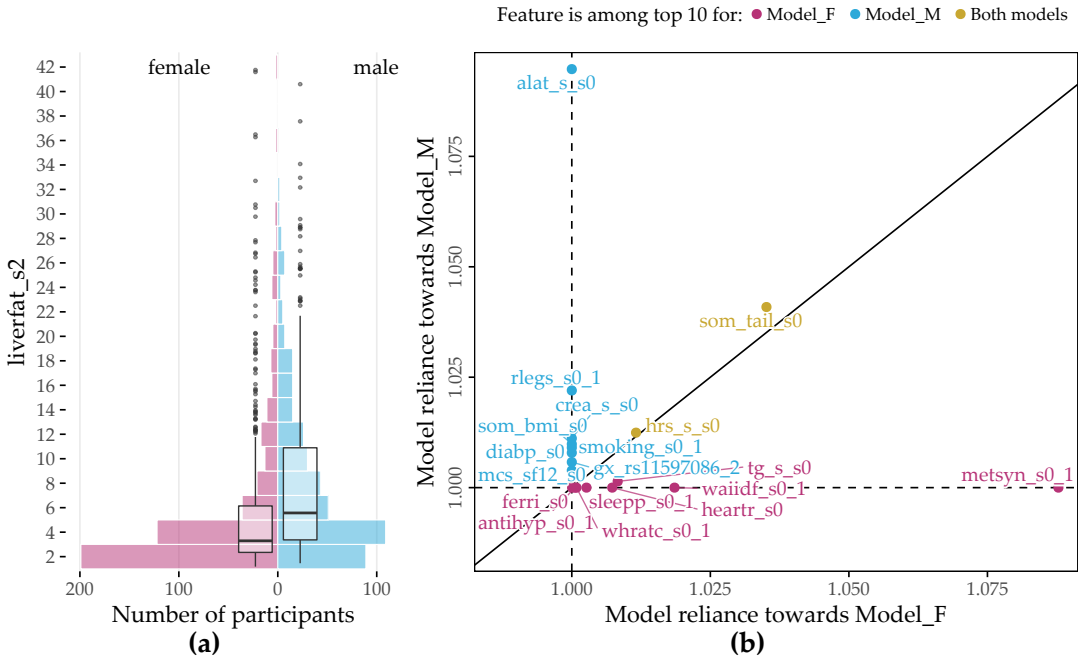


Figure 9.6: **Distribution of target variable and variable importance (LT 5)**. (a) Distribution of liverfat_s2 for the subset of female and male SHIP participants. (b) Each scatterplot shows the model reliance (MR) score for each predictor for the best model for each of the subpopulations of female (x-axis) and male (y-axis) patients and each learning task; see Figure 9.1. Variables among the top 10 highest-ranked variables by MR in the F_model (red-violet), M_model (blue), or both models (yellow) are highlighted. alat_s_s0: alanin aminotransferase (ALAT) concentration; antihyp_s0_1: antihypertensive medication; crea_s_s0: serum creatinine concentration; diabp_s0: diastolic blood pressure; ferri_s0: ferritin concentration; gx_rs11597086_2: genetic marker associated with ALAT concentration [300]; heartr_s0: heart rate; hrs_s_s0: serum uric acid concentration; mcs_sf12_s0: SF-12 Physical and Mental Health Summary Scale [38]; metsyn_s0_1: metabolic syndrome; rlegs_s0_1: restless legs syndrome; sleepp_s0_1: sleep problems; smoking_s0_1: ex-smoker; som_bmi_s0: body mass index; som_tail_s0: waist circumference; tg_s_s0: serum triglycerides concentration; waiidf_s0_1: increased waist circumference; whratc_s0_1: waist to hip ratio.

9.6 Conclusion

We have presented a workflow to juxtapose the most important predictors between two a priori defined subpopulations based on black-box models. We have adapted model reliance to estimate a feature's attribution regarding the model and to investigate subpopulation-specific differences in this respect. Compared to the workflow in Chapter 8, our goal was not to find a parsimonious model; hence we did not perform feature selection.

Our goals are related to *causal inference*, which aims to measure the true, *unconfounded* effect between variables. Current efforts include building methods to detect causal relationships in observational data [209, 239], opposed to a randomized clinical trial (RCT) in clinical research. RCTs are considered the gold standard for inferring causal effects of treatment [116]. In an RCT, individual patients of a patient population are randomly assigned to one of two subgroups: a *treatment* subgroup and a *control* subgroup. Only the former receives the treatment. Randomization of subgroup membership serves to minimize the effects of potential confounders

and selection bias. The two subgroups are as similar as possible before the intervention so that it is possible to calculate the *average treatment effect* to quantify the true causal efficacy of the treatment [122]. Translated to our application, an appropriate causality question is: *What is the effect of gender on tinnitus severity and depression?* However, our goal was different: we wanted to exploratively examine similarities and differences between subpopulations regarding factors that are predictive of a response of interest. Thus, we were interested not only in the relationship of gender on the response, but also in differences between the female and male subpopulations in the predictability of other variables on the response. With the generation of new hypotheses as the main goal in mind, our focus is on providing exploratory methods for comparing differences between predefined subpopulations.

Part IV

CONCLUSION

Chapter 10

Summary and Future Work

Data-driven machine learning solutions can complement the traditionally *hypothesis-driven* workflows of medical research by discovering characteristic *subpopulations* of patients and study participants. Knowledge about characteristic subpopulations can be the starting point for further investigation, e.g., identifying (long-term) risk factors, determining differences in treatment response, and building robust statistical models that explain cause-effect relationships for a medical condition of interest.

This thesis has proposed solutions to assist expert-driven subpopulation discovery in high-dimensional timestamped medical data. Section 10.1 reviews whether we have addressed our core research question and the three challenges presented in Chapter 1. Finally, we discuss directions for future work in Section 10.2.

10.1 Summary

Our solutions for subpopulation discovery in Part I have tackled the challenge:

GOAL1: *Comprehensibility and distinctiveness of subpopulations*

We have dealt with this challenge by proposing the three workflows presented in Chapters 3, 4, and 5.

Chapter 3 has presented a workflow and an interactive application for subpopulation discovery in cohort data. By leveraging classification rules, our workflow enables building self-explaining and concise descriptions of subpopulations with distributions regarding the target variable. Our application *Interactive Medical Miner* allows for interactive expert-driven subpopulation discovery with features to drill-down on the derived models and explore subpopulations worthy of further investigation. As a proof of concept, we have validated our workflow on a SHIP dataset, focusing on hepatic steatosis ("fatty liver") as the target variable. We have confirmed variables and value ranges shown to be associated with the target variable, including obesity, age, sex, high serum concentrations of the liver enzyme Gamma-glutamyltransferase (GGT), and genetic markers.

Chapter 4 has extended this workflow by tackling the problem of redundancy in large rule sets, introducing an algorithm that extracts a small number of representative classification rules. These so-called *proxy rules* minimize instance overlap across rule groups, thus covering different subpopulations. We have demonstrated that our algorithm delivers *more distinct* rules compared to the baselines on SHIP data samples with hepatic steatosis and goiter as target

variables, respectively. Additionally to the variables found in Chapter 3, the conditions of proxy rules involve hypertension and low physical health for hepatic steatosis, and high intima-media thickness and angiotensin II receptor blocker intake for goiter.

The solutions in Chapters 3 and 4 assume the availability of a target variable, which in many medical applications is unknown or too costly to obtain. Hence, Chapter 5 has proposed a workflow for unsupervised subpopulation discovery, visualization, and interactive exploration in the absence of a ground truth. This workflow (i) exploits a clustering algorithm that automatically determines an appropriate number of clusters, (ii) provides visualization techniques that show essential characteristics of high-dimensional clusters compactly, and (iii) serves medical experts with a web application to look into and juxtapose the found subpopulations further, including treatment effect indicators. We have showcased our workflow's efficacy by identifying and juxtaposing four distinct tinnitus patient *phenotypes* in the CHA dataset.

While the aforementioned solutions are designed primarily for static data, Part II of this thesis has addressed the exploitation of temporal information in medical data, translated to the challenge:

GOAL2: *Exploitation of time*

We have presented our solutions to this challenge in Chapters 6 and 7.

Chapter 6 has proposed a framework to extract evolution features from timestamped medical data, i.e., tiny streams with up to five time points that are months or years apart (which is often the case in population studies). These change descriptors quantify the study participants' change over time. We have shown that augmenting the original feature space with evolution features improves classification performance. Using a SHIP data sample, we have shown that somatic changes and changes of cluster quality indices over time are associated with hepatic steatosis.

Chapter 7 has focused on data with many recordings over a short time period and proposes a clustering approach to build representations from new similarity measures applicable to raw multivariate (sensor) timeseries. We have validated our approach by identifying plantar pressure patterns in patients with diabetic foot syndrome and healthy volunteers.

Part III has presented solutions for models that cannot be interpreted intrinsically. Because medical decisions have serious consequences, more accurate models are increasingly preferred in research as they suggest a higher degree of confidence in their output. However, post-modeling methods are needed to translate the findings of these black-box models into an expert-understandable form, leading to the challenge:

GOAL3: *Post-hoc interpretation of complex black-box models*

Chapter 8 has proposed an end-to-end data analysis workflow with steps for data augmentation, modeling, interleaving model training with feature elimination, and post-hoc analysis of the trained models. Our workflow yields (for any type of model) statistics and visualizations representing global feature importance, instance-individual feature importance, and subpopulation-specific feature importance. We have validated our workflow on three modeling tasks: (i) tinnitus-related distress after treatment in tinnitus patients, (ii) depression at baseline in tinnitus patients, and (iii) rupture risk in intracranial aneurysms.

Finally, Chapter 9 has presented a solution to examine how pre-defined subpopulations differ concerning their most predictive characteristics. We have derived a post-hoc interpretation measure to assess differences in the predictors' associations between two subpopulations. We have validated our solution by finding gender and sex differences (subpopulations of female and male patients) for the target variables tinnitus-related distress, depression and liver fat concentration.

10.2 Future Work

Causal subpopulation discovery. While our machine learning solutions can initiate the generation of new hypotheses, they are limited to discovering mere *associations* between variables, not *cause-effect relationships*. For example, we have shown that depression strongly correlates with distress in tinnitus patients (recall Chapters 5 and 8). The *direction* of the relationship cannot be inferred – does tinnitus distress cause depression or vice versa? Or can the correlation instead be explained by a *confounder*, i.e., a "lurking" third variable, such as low physical health or socioeconomic status? Pearl and Mackenzie [209] give an example of the *Simpson's paradox*, a phenomenon where traditional correlation analysis even contradicts true cause-effect relationships. When considering only pairwise relationships, the data suggested that people who exercise more had increased cholesterol levels. However, this relationship was masked by age. Older participants exhibited higher activity levels in the study population. In an age-stratified analysis, it became evident that more exercise leads to a decrease in cholesterol levels.

In causal inference, relationships between variables are modeled as a directed acyclic graph (DAG) where the nodes represent the variables and the edges the direction of causality [122, 209]. Schölkopf [239] described potential links between machine learning and causal inference. Yu et al. [299] reviewed causal feature selection methods, which assume that the optimal feature subset for a target variable is equal to its *Markov boundary*. The Markov boundary comprises nodes with outgoing (*parents*) or incoming (*children*) edges to the target and other parents of child nodes (*spouses*) in the DAG (Figure 10.1). The assumption is that by conditioning on a target variable's Markov boundary, all other variables in the DAG become independent of the target [299]. Hence, a possible direction for future work is to investigate the potential of combining our methods with causal inference, for example, by leveraging causal feature selection for predictive modeling. Here, we are interested in determining whether features in the Markov boundary are sufficient to predict the response and how causal approaches compare against their non-causal counterparts regarding classification performance.

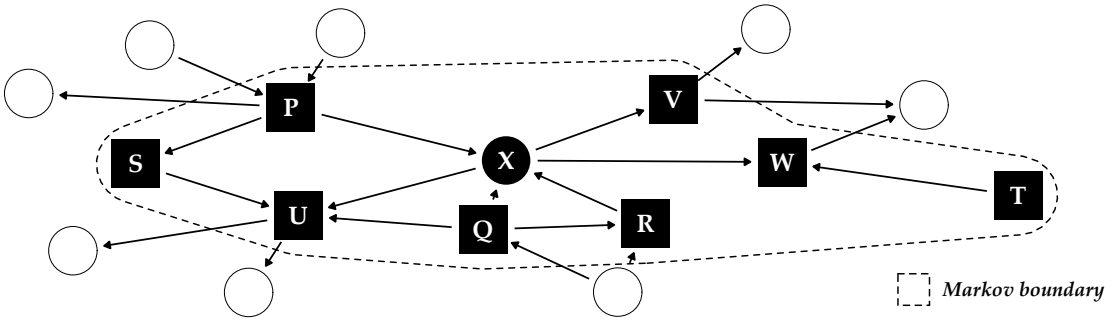


Figure 10.1: **Illustrative example of a directed acyclic graph and Markov boundary for the target X.** The other labeled variables P-W constitute the Markov boundary of X, consisting of its parents, children and any additional parents of its children ("spouses").

Domain adaption. For the time being, the findings on the datasets investigated in this thesis cannot be replicated on other datasets. For example, we expect that the models of our evolution feature framework (Chapter 6) built for SHIP will perform poorly when applied to data samples from other population studies, such as MONICA [143], KORA [136], or the Rotterdam Study [135]. *Domain adaption* deals with transferring knowledge from the *source* to the *target domain*. Lemberger and Panico [178] summarized four major challenges in domain adaption:

- *Prior shift* refers to a difference between the source (A) and the target (B) domain in the target variable's distribution. For example, in this situation, the liver fat concentration in A is (on average) higher than in B. This could be due to differences in sampling, e.g., when for A, a higher percentage of subjects with pre-existing conditions were drawn.
- *Covariate shift* refers to a difference in the predictors' distributions in A and B, while the relationships between the predictors and the target variable are assumed to be the same. For example, subjects in A are older than subjects in B, while at the same time, a high body mass index is predictive in both domains.
- *Concept shift* refers to a difference in the relationship between the predictors and the target variable. For example, while in A, smoking is observed in subjects with increased liver fat concentration, smokers in B generally exhibit lower liver fat values than non-smokers.
- *Subspace mapping* refers to the situation where the feature sets in A and B are different. For example, hepatic steatosis is determined by ultrasound in A and by MRT in B. Besides, different questionnaires were used in A and B to assess life quality, and neither of the questionnaires in A was used in B and vice versa.

Future efforts include leveraging methods from domain adaption, for example, to transfer the tinnitus phenotypes from Chapter 5 detected in CHA to patient data of other tinnitus centers with different baseline characteristics.

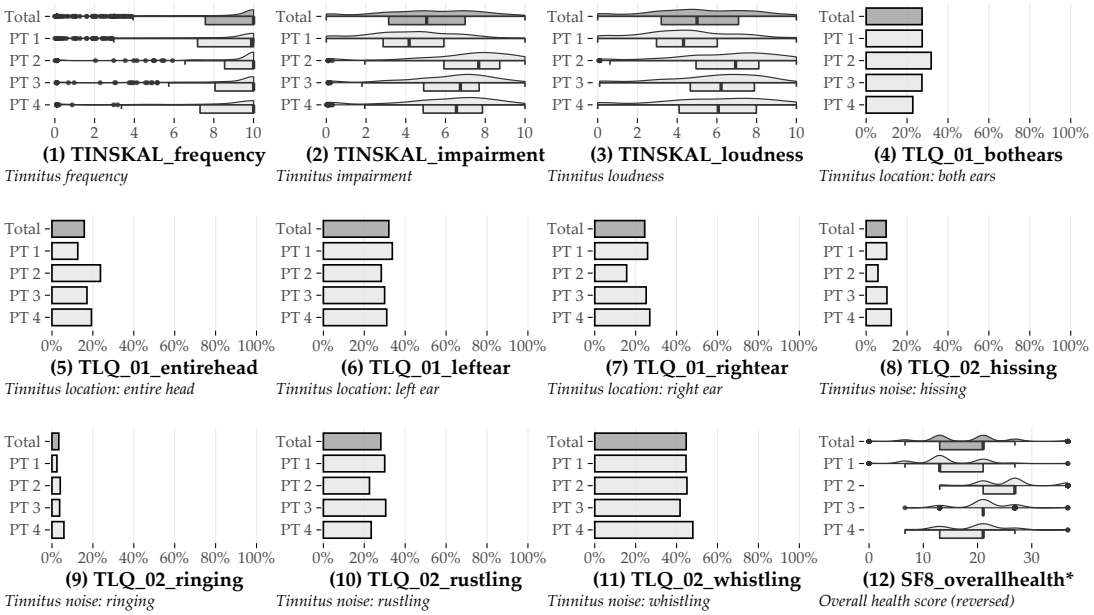
Parsimonious and cost-aware learning. Our methods disregard the fact that features are not always available for free. Feature acquisition in medical research is associated with certain *costs*, divided into the financial expense and the health burden for the patient undergoing a potentially painful and invasive examination [149]. For example, a biopsy usually has a higher diagnostic value than a simple blood test but is also more expensive and physically and mentally stressful for the patient. Yu et al. [298] perform feature selection under a budget to balance model performance and feature acquisition cost. Consequently, their models favor cheaper features like demographics and questionnaire responses over costly ones requiring more expensive or strenuous physical examinations or laboratory tests. Recall Chapters 3 and 4, where both somatic and ultrasound variables appeared in the classification rules describing subpopulations with increased liver fat concentration. Our methods do not distinguish between variables of simple body measurements (e.g., to obtain a participant's waist circumference or body weight) and elaborate imaging procedures. Future research includes implementing mechanisms to enable cost-aware learning.

Appendix A

Overview of Variables Selected for Phenotyping

For each of the 64 variables selected for phenotyping in Chapter 5, phenotype-stratified distributions and a brief description are shown below. Half-violin/half-boxplot geometries represent numerical variables; bar graphs represent categorical variables. The order of the variables is analogous to the clockwise arrangement in Figures 5.2 and 5.3. Variables with an asterisk at the end of their name were reversed to ensure higher scores consistently represent higher health burden across all variables.

The prefix of a variable's name denotes the respective questionnaire: ACSA: Anamnestic Comparative Self-Assessment [26]; ADSL: General Depression Scale [222, 119]; BI: Berlin Complaint Inventory [139]; BSF: Berlin Mood Questionnaire [138]; ISR: ICD-10 Symptom Rating [267]; PHQK: (Short-form) Patient Health Questionnaire [252]; PSQ: Perceived Stress Questionnaire [80]; SES: Pain Perception Scale [90]; SF8: Short Form 8 Health Survey [37]; SOZK: A sociodemographics questionnaire [36]; SWOP: Self-Efficacy- Optimism-Pessimism Scale questionnaire [240]; TINSKAL: Visual analog scales; TLQ: Tinnitus Localization and Quality questionnaire [97]; TQ: Tinnitus Questionnaire (German version) [98].



(1) TINSKAL_frequency
Tinnitus frequency

(2) TINSKAL_impairment
Tinnitus impairment

(3) TINSKAL_loudness
Tinnitus loudness

(4) TLQ_01_bothears
Tinnitus location: both ears

(5) TLQ_01_entirehead
Tinnitus location: entire head

(6) TLQ_01_leftear
Tinnitus location: left ear

(7) TLQ_01_rightear
Tinnitus location: right ear

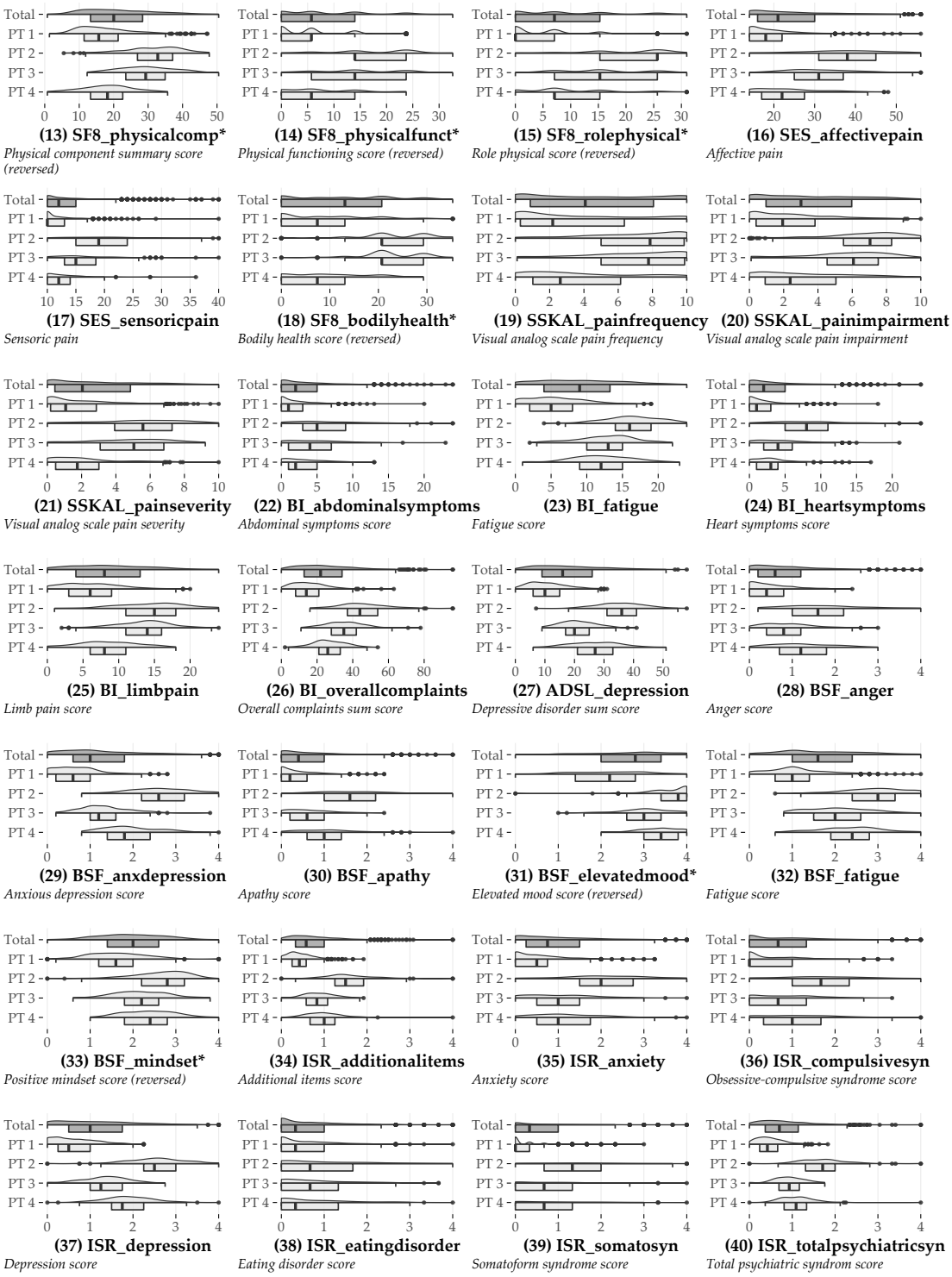
(8) TLQ_02_hissing
Tinnitus noise: hissing

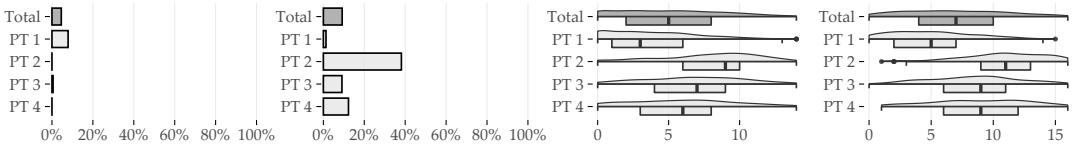
(9) TLQ_02_ringing
Tinnitus noise: ringing

(10) TLQ_02_rustling
Tinnitus noise: rustling

(11) TLQ_02_whistling
Tinnitus noise: whistling

(12) SF8_overallhealth*
Overall health score (reversed)


Total
PT 1
PT 2
PT 3
PT 4
(13) SF8_physicalcomp*
Physical component summary score (reversed)
(14) SF8_physicalfunct*
Physical functioning score (reversed)
(15) SF8_rolephysical*
Role physical score (reversed)
(16) SES_affectivepain
Affective pain
(17) SES_sensoricpain
Sensoric pain
(18) SF8_bodilyhealth*
Bodily health score (reversed)
(19) SSKAL_painfrequency
Visual analog scale pain frequency
(20) SSKAL_painimpairment
Visual analog scale pain impairment
(21) SSKAL_painseverity
Visual analog scale pain severity
(22) BI_abdominalsymptoms
Abdominal symptoms score
(23) BI_fatigue
Fatigue score
(24) BI_heartsymptoms
Heart symptoms score
(25) BI_limbpain
Limb pain score
(26) BI_overallcomplaints
Overall complaints sum score
(27) ADSL_depression
Depressive disorder sum score
(28) BSF_anger
Anger score
(29) BSF_anxdepression
Anxious depression score
(30) BSF_apathy
Apathy score
(31) BSF_elevatedmood*
Elevated mood score (reversed)
(32) BSF_fatigue
Fatigue score
(33) BSF_mindset*
Positive mindset score (reversed)
(34) ISR_additionalitems
Additional items score
(35) ISR_anxiety
Anxiety score
(36) ISR_compulsivesyn
Obsessive-compulsive syndrome score
(37) ISR_depression
Depression score
(38) ISR_eatingdisorder
Eating disorder score
(39) ISR_somatosyn
Somatoform syndrome score
(40) ISR_totalpsychiatricsyn
Total psychiatric syndrom score

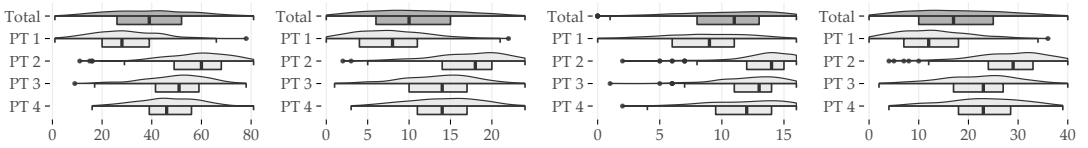


(41) PHQK_depression
Presence of depression

(42) PHQK_panicsyn
Presence of panic syndrome

(43) TQ_auditoryperceptdiff
Auditory perceptual difficulties score

(44) TQ_cognitivedistress
Cognitive distress score

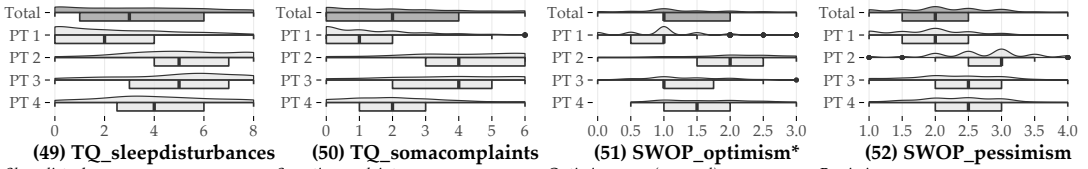


(45) TQ_distress
Total tinnitus distress score

(46) TQ_emodistress
Emotional distress score

(47) TQ_intrusiveness
Intrusiveness score

(48) TQ_psychodistress
Psychological distress score

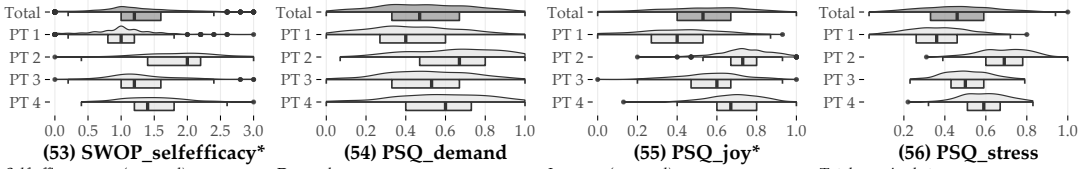


(49) TQ_sleepdisturbances
Sleep disturbances score

(50) TQ_somacomplaints
Somatic complaints score

(51) SWOP_optimism*
Optimism score (reversed)

(52) SWOP_pessimism
Pessimism score

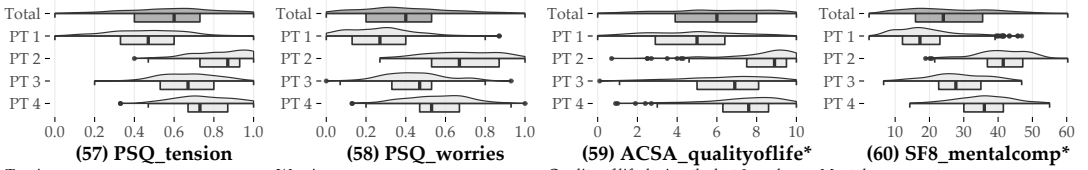


(53) SWOP_selfefficacy*
Self-efficacy score (reversed)

(54) PSQ_demand
Demand score

(55) PSQ_joy*
Joy score (reversed)

(56) PSQ_stress
Total perceived stress sum score

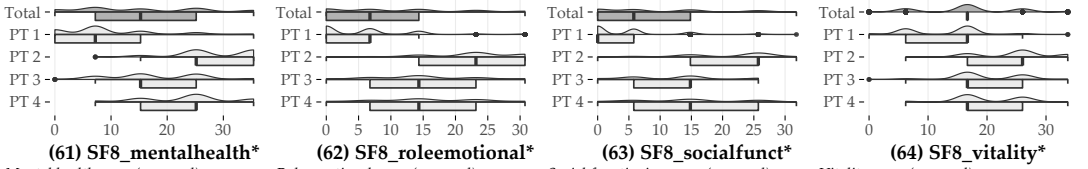


(57) PSQ_tension
Tension score

(58) PSQ_worries
Worries score

(59) ACSA_qualityoflife*
Quality of life during the last 2 weeks (reversed)

(60) SF8_mentalcomp*
Mental component summary score (reversed)



(61) SF8_mentalhealth*
Mental health score (reversed)

(62) SF8_roleemotional*
Role emotional score (reversed)

(63) SF8_socialfunct*
Social functioning score (reversed)

(64) SF8_vitality*
Vitality score (reversed)

Bibliography

[1] Caroline A. Abbott et al. "Innovative intelligent insole system reduces diabetic foot ulcer recurrence at plantar sites: a prospective, randomised, proof-of-concept study". In: *The Lancet Digital Health* 1.6 (2019), e308–e318. DOI: 10.1016/S2589-7500(19)30128-1. URL: https://www.sciencedirect.com/science/article/pii/S2589750019301281.

[2] 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.

[3] Charu C. Aggarwal. "Data Classification: Algorithms and Applications". In: CRC Press, 2014. Chap. A Survey of Stream Classification Algorithms, pp. 245–273.

[4] Shiva Alemzadeh et al. "Subpopulation Discovery and Validation in Epidemiological Data". In: *EuroVis Workshop on Visual Analytics (EuroVA)*. 2017. ISBN: 978-3-03868-042-0. DOI: 10.2312/eurova.20171118. URL: https://doi.org/10.2312/eurova.20171118.

[5] David Alvarez-Melis and Tommi S. Jaakkola. *On the Robustness of Interpretability Methods*. 2018. eprint: https://arxiv.org/abs/1806.08049.

[6] Luca Antiga et al. "An image-based modeling framework for patient-specific computational hemodynamics". In: *Medical & Biological Engineering & Computing* 46.11 (2008), pp. 1097–1112. DOI: 10.1007/s11517-008-0420-1.

[7] Catiele Antunes, Mohammadreza Azadfard, and Mohit Gupta. *Fatty Liver*. StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/books/NBK441992. 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK441992/.

[8] David G. Armstrong et al. "Is there a critical level of plantar foot pressure to identify patients at risk for neuropathic foot ulceration?" In: *The Journal of Foot and Ankle Surgery* 37.4 (1998), pp. 303–307. DOI: 10.1016/S1067-2516(98)80066-5.

[9] M. L. Arts et al. "Data-driven directions for effective footwear provision for the high-risk diabetic foot". In: *Diabetic Medicine* 32.6 (2015), pp. 790–797. DOI: 10.1111/dme.12741. URL: https://www.ncbi.nlm.nih.gov/pubmed/25763659.

[10] Maya Asher, Anu Asnaani, and Idan M. Aderka. "Gender differences in social anxiety disorder: A review". In: *Clinical Psychology Review* 56 (2017), pp. 1–12. DOI: https://doi.org/10.1016/j.cpr.2017.05.004.

[11] Martin Atzmüller. "Subgroup discovery". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.1 (2015), pp. 35–49. DOI: https://doi.org/10.1002/widm.1144.

[12] Martin Atzmüller and Florian Lemmerich. "VIKAMINE – Open-Source Subgroup Discovery, Pattern Mining, and Analytics". In: *Machine Learning and Knowledge Discovery in Databases*. 2012, pp. 842–845. DOI: 10.1007/978-3-642-33486-3_60.

[13] Martin Atzmüller and Frank Puppe. "SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery". In: *Machine Learning and Knowledge Discovery in Databases*. 2006, pp. 6–17. DOI: 10.1007/11871637_6.

[14] 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: *Journal of Clinical Epidemiology* 66.4 (2013), pp. 398–407. DOI: 10.1016/j.jclinepi.2012.11.008.

[15] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7 (2015), pp. 1–46. DOI: 10.1371/journal.pone.0130140.

[16] Francesco Bagattini et al. "A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records". In: *BMC Medical Informatics and Decision Making* 19.1 (2019), pp. 1–20. DOI: 10.1186/s12911-018-0717-4.

[17] David Baguley, Don McFerran, and Deborah Hall. "Tinnitus". In: *The Lancet* 382.9904 (2013), pp. 1600–1607. DOI: 10.1016/S0140-6736(13)60142-7.

[18] 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. DOI: 10.3171/2011.11.JNS11311.

[19] Ruth Barn et al. "Predictors of Barefoot Plantar Pressure during Walking in Patients with Diabetes, Peripheral Neuropathy and a History of Ulceration". In: *PLOS ONE* 10.2 (2015), pp. 1–12. DOI: 10.1371/journal.pone.0117443.

[20] Sebastian E. Baumeister et al. "Impact of Fatty Liver Disease on Health Care Utilization and Costs in a General Population: A 5-Year Observation". In: *Gastroenterology* 134.1 (2008), pp. 85–94. DOI: 10.1053/j.gastro.2007.10.024.

[21] Giorgio Bedogni et al. "The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population". In: *BMC Gastroenterology* 6.33 (2006), pp. 1–7. DOI: 10.1186/1471-230X-6-33.

[22] Stefano Bellentani et al. "Prevalence of and Risk Factors for Hepatic Steatosis in Northern Italy". In: *BMC Gastroenterology* 132.2 (2000), pp. 112–117. DOI: 10.7326/0003-4819-132-2-200001180-00004.

[23] Craig J. Bennetts et al. "Clustering and Classification of Regional Peak Plantar Pressures of Diabetic Feet". In: *Journal of Biomechanics* 46.1 (2013), pp. 19–25. DOI: 10.1016/j.jbiomech.2012.09.007.

[24] 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 (IJCARS)* 13.1 (2018), pp. 83–93. DOI: 10.1007/s11548-017-1643-0.

[25] 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.

[26] 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.

[27] Juan-Jose Beunza et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)". In: *Journal of Biomedical Informatics* 97.103257 (2019), pp. 1–6. DOI: 10.1016/j.jbi.2019.103257.

[28] 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.

[29] Bernd Bischl et al. "mlr: Machine Learning in R". In: *Journal of Machine Learning Research* 17.170 (2016), pp. 1–5. URL: https://jmlr.org/papers/v17/15-066.html.

[30] Amelia K. Boehme, Charles Esenwa, and Mitchell S. V. Elkind. "Stroke Risk Factors, Genetics, and Prevention". In: *Circulation Research* 120.3 (2017), pp. 472–495. DOI: 10.1161/CIRCRESAHA.116.308398.

[31] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classifiers". In: *Workshop on Computational Learning Theory*. 1992, pp. 144–152. DOI: 10.1145/130385.130401.

[32] A. J. Boulton et al. "The global burden of diabetic foot disease". In: *The Lancet* 366.9498 (2005), pp. 1719–1724. DOI: 10.1016/S0140-6736(05)67698-2. URL: https://www.ncbi.nlm.nih.gov/pubmed/16291066.

[33] E. J. Boyko, A. D. Seelig, and J. H. Ahroni. "Limb- and Person-Level Risk Factors for Lower-Limb Amputation in the Prospective Seattle Diabetic Foot Study ". In: *Diabetes Care* 41.4 (2018), pp. 891–898. DOI: 10.2337/dc17-2210. URL: https://www.ncbi.nlm.nih.gov/pubmed/29439130.

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

[35] Leo Breiman. "Random Forests". In: *Machine learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[36] Petra Brueggemann et al. "Impact of Multiple Factors on the Degree of Tinnitus Distress". In: *Frontiers in Human Neuroscience* 10.341 (2016), pp. 1–11. DOI: 10.3389/fnhum.2016.00341.

[37] M. Bullinger and M. Morfeld. "Der SF-36 Health Survey". In: *Gesundheitsökonomische Evaluationen*. Springer, 2008, pp. 387–402. DOI: 10.1007/978-3-540-49559-8_15.

[38] Monika Bullinger. "German translation and psychometric testing of the SF-36 Health Survey: Preliminary results from the IQOLA project". In: *Social Science & Medicine* 41.10 (1995), pp. 1359–1366. DOI: 10.1016/0277-9536(95)00115-N.

[39] Sicco A. Bus. "Priorities in offloading the diabetic foot". In: *Diabetes/Metabolism Research and Reviews* 28.S1 (2012), pp. 54–59. DOI: 10.1002/dmrr.2240.

[40] Josip Car et al. "Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom". In: *BMC Medicine* 17.143 (2019). DOI: 10.1186/s12916-019-1382-x.

[41] 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. DOI: 10.1007/s00500-010-0670-3.

[42] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". In: *Electronics* 8.8 (2019), pp. 1–34. DOI: 10.3390/electronics8080832.

[43] Peter R. Cavanagh and Sicco A. Bus. "Off-loading the diabetic foot for ulcer prevention and healing". In: *Journal of Vascular Surgery* 52.3 (2010), 37S–43S. DOI: 10.1016/j.jvs.2010.06.007.

[44] 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. DOI: 10.3174/ajnr.A2419.

[45] Christopher R. Cederroth, Silvano Gallus, Deborah A. Hall, et al. "Towards an Understanding of Tinnitus Heterogeneity". In: *Frontiers in Aging Neuroscience* 11.53 (2019), pp. 1–7. DOI: 10.3389/fnagi.2019.00053.

[46] 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.

[47] Nitesh V. Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16.1 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[48] Jonathan H. Chen and Steven M. Asch. "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations". In: *The New England Journal of Medicine* 376.26 (2017), pp. 2507–2509. DOI: 10.1056/NEJMp1702071.

[49] Tianqi Chen et al. *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.2. 2019. URL: https://CRAN.R-project.org/package=xgboost.

[50] Flavia Agata Cimini, Ilaria Barchetta, Gea Ciccarelli, et al. "Adipose tissue remodelling in obese subjects is a determinant of presence and severity of fatty liver disease". In: *Diabetes/Metabolism Research and Reviews* 37.1 (2021), pp. 1–13. DOI: 10.1038/s41467-020-16540-x.

[51] Peter Clark and Tim Niblett. "The CN2 induction algorithm". In: *Machine Learning* 3.4 (1989), pp. 261–283. DOI: 10.1007/BF00116835.

[52] William S. Cleveland. "Robust Locally Weighted Regression and Smoothing Scatterplots". In: *Journal of the American Statistical Association* 74.368 (1979), pp. 829–836. DOI: 10.1080/01621459.1979.10481038.

[53] Jordana B. Cohen, Sarah J. Schrauben, Lei Zhao, et al. "Clinical Phenogroups in Heart Failure With Preserved Ejection Fraction: Detailed Phenotypes, Prognosis, and Response to Spironolactone". In: *Heart Failure* 8.3 (2020), pp. 172–184. DOI: 10.1016/j.jchf.2019.09.009.

[54] William W. Cohen. "Fast effective rule induction". In: *International Conference on Machine Learning*. 1995, pp. 115–123. DOI: 10.1016/B978-1-55860-377-6.50023-2.

[55] Carlo Combi, Elpida Keravnou-Papailiou, and Yuval Shahar. *Temporal Information Systems in Medicine*. Springer, 2010. DOI: 10.1007/978-1-4419-6543-1.

[56] Alberto Corvo et al. "Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology". In: *Transactions on Visualization and Computer Graphics (TVCG)* (2020), pp. 1–18. DOI: 10.1109/TVCG.2020.2990336.

[57] Rutger Dahmen et al. "Therapeutic Footwear for the Neuropathic Foot: An algorithm". In: *Diabetes Care* 24.4 (2001), pp. 705–709. DOI: 10.2337/diacare.24.4.705.

[58] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997. DOI: 10.1017/CBO9780511802843.

[59] Anneleen De Cock et al. "A functional foot type classification with cluster analysis based on plantar pressure distribution during jogging". In: *Gait & Posture* 23.3 (2006), pp. 339–347. DOI: 10.1016/j.gaitpost.2005.04.011.

[60] Kevin Deschamps et al. "Classification of Forefoot Plantar Pressure Distribution in Persons with Diabetes: A Novel Perspective for the Mechanical Management of Diabetic Foot?" In: *PLOS ONE* 8.11 (2013), pp. 1–10. DOI: 10.1371/journal.pone.0079924.

[61] 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 (IJCARS)* 13.11 (2018), pp. 1767–1779. DOI: 10.1007/s11548-018-1837-0.

[62] Sujan Dhar et al. “Morphology Parameters for Intracranial Aneurysm Rupture Risk Assessment”. In: *Neurosurgery* 63.2 (2008), pp. 185–197. DOI: 10.1227/01.neu.0000316847.64140.81.

[63] Lee R. Dice. “Measures of the Amount of Ecologic Association Between Species”. In: *Ecology* 26.3 (1945), pp. 297–302. DOI: https://doi.org/10.2307/1932409.

[64] Thomas J. DiCiccio and Bradley Efron. “Bootstrap confidence intervals”. In: *Statistical Science* 11.3 (1996), pp. 189–212. DOI: 10.1214/ss/1032280214.

[65] Beiying Ding and Robert Gentleman. “Classification Using Generalized Partial Least Squares”. In: *Journal of Computational and Graphical Statistics* 14.2 (2005), pp. 280–298. DOI: 10.1198/106186005X47697.

[66] Robert A. Dobie. “Depression and tinnitus”. In: *Otolaryngologic Clinics of North America* 36.2 (2003), pp. 383–388. DOI: 10.1016/s0030-6665(02)00168-8.

[67] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. eprint: https://arxiv.org/abs/1702.08608.

[68] 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*. 2010, pp. 192–196. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041306/.

[69] Joann G. Elmore et al. *Jekel’s Epidemiology, Biostatistics and Preventive Medicine*. Elsevier Health Sciences, 2020.

[70] Alev Akyol Erikci et al. “The effect of subclinical hypothyroidism on platelet parameters”. In: *Hematology* 14.2 (2009), pp. 115–117. DOI: 10.1179/102453309X385124.

[71] Soly I. Erlandsson and Kajsa-Mia Holgers. “The impact of perceived tinnitus severity on health-related quality of life with aspects of gender”. In: *Noise and Health* 3.10 (2001), pp. 39–51. URL: http://www.noiseandhealth.org/article.asp?issn=1463-1741;year=2001;volume=3;issue=10;spage=39;epage=51;aulast=Erlandsson.

[72] Martin Ester et al. “A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Knowledge Discovery and Data Mining (KDD)*. Vol. 96. 34. 1996, pp. 226–231.

[73] A. S. 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. DOI: 10.1111/j.1742-1241.2007.01534.x. URL: https://www.ncbi.nlm.nih.gov/pubmed/17935551.

[74] Hassan Ismail Fawaz et al. “Deep learning for time series classification: a review”. In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963. DOI: 10.1007/s10618-019-00619-1.

[75] Usama M. Fayyad and Keki B. Irani. “Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 1993, pp. 1022–1027.

[76] Serafino Fazio et al. “Effects of Thyroid Hormone on the Cardiovascular System”. In: *Recent Progress in Hormone Research* 59.1 (2004), pp. 31–50. DOI: 10.1210/rp.59.1.31.

[77] Malindu Fernando, Scott Wearing, and Robert Crowther. “Handbook of Human Motion”. In: Springer, 2018. Chap. The Importance of Foot Pressure in Diabetes, pp. 759–787. DOI: 10.1007/978-3-319-14418-4_39.

[78] Hazel Finney, David J. Newman, and Christopher P. Price. “Adult Reference Ranges for Serum Cystatin C, Creatinine and Predicted Creatinine Clearance”. In: *Annals of Clinical Biochemistry* 37.1 (2000), pp. 49–59. DOI: 10.1258/0004563001901524.

[79] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously". In: *Journal of Machine Learning Research* 20.177 (2019). URL: http://jmlr.org/papers/v20/18-760.html, pp. 1–81.

[80] 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. DOI: 10.1097/01.psy.0000151491.80178.78.

[81] Robert L. Folmer et al. "Tinnitus severity, loudness, and depression". In: *Otolaryngology—Head and Neck Surgery* 121.1 (1999), pp. 48–51. DOI: 10.1016/S0194-5998(99)70123-3.

[82] 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.

[83] Pascal Friederich et al. "Scientific intuition inspired by machine learning generated hypotheses". In: *Machine Learning: Science and Technology* (2021). DOI: 10.1088/2632-2153/abda08.

[84] 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. DOI: 10.18637/jss.v033.i01. URL: http://www.jstatsoft.org/v33/i01/.

[85] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine". In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[86] Robert G. Frykberg, Thomas Zgonis, David G. Armstrong, et al. "Diabetic foot disorders. A clinical practice guideline (2006 revision)". In: *The Journal of Foot and Ankle Surgery* 45.5 (2006), pp. 1–66. DOI: 10.1016/S1067-2516(07)60001-5.

[87] Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of Rule Learning*. Springer Science & Business Media, 2012. DOI: 10.1007/978-3-540-75197-7.

[88] Dragan Gamberger and Nada Lavrac. "Expert-Guided Subgroup Discovery: Methodology and Application". In: *Journal of Artificial Intelligence Research* 17 (2002), pp. 501–527. DOI: 10.1613/jair.1089.

[89] Ning Gao et al. "Carotid intima-media thickness in patients with subclinical hypothyroidism: a meta-analysis". In: *Atherosclerosis* 227.1 (2013), pp. 18–25. DOI: 10.1016/j.atherosclerosis.2012.10.070.

[90] 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.

[91] Eleni Genitsaridi et al. "A Review and a Framework of Variables for Defining and Characterizing Tinnitus Subphenotypes". In: *Brain Sciences* 10.12 (2020), pp. 1–21. DOI: 10.3390/brainsci10120938.

[92] Ulf Gerhardt, Ruediger Breitschwerdt, and Oliver Thomas. "mHealth Engineering: A Technology Review". In: *Journal of Information Technology Theory and Application* 19.3 (2018). URL: https://aisel.aisnet.org/jitta/vol19/iss3/5, pp. 82–117. URL: https://aisel.aisnet.org/jitta/vol19/iss3/5.

[93] Claudia Giacomozzi and Francesco Martelli. "Peak pressure curve: an effective parameter for early detection of foot functional impairments in diabetic patients". In: *Gait & Posture* 23.4 (2006), pp. 464–470. DOI: 10.1016/j.gaitpost.2005.06.006.

[94] David Gilbert. *JFreeChart (Free Java class library for creating charts)*. URL: https://www.jfree.org/jfreechart/. 2005-2021. URL: http://www.jfree.org/jfreechart/.

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

[96] Norval D. Glenn. *Cohort analysis*. Second edition. Sage, 2005. DOI: 10 . 4135 / 9781412983662.

[97] Gerhard Goebel and Wolfgang Hiller. "Psychische Beschwerden bei chronischem Tinnitus: Erprobung und Evaluation des Tinnitus-Fragebogens (TF)". In: *Verhaltenstherapie* 2.1 (1992), pp. 13–22. DOI: 10.1159/000258202.

[98] Gerhard Goebel and Wolfgang Hiller. *Tinnitus-Fragebogen (TF). Ein Instrument zur Erfassung von Belastung und Schweregrad bei Tinnitus*. Hogrefe, 1998.

[99] 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.

[100] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[101] J. C. Gower. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis". In: *Biometrika* 53.3/4 (1966), pp. 325–338. DOI: 10.2307/2333639.

[102] E. W. Gregg et al. "Changes in Diabetes-Related Complications in the United States, 1990–2010". In: *The New England Journal of Medicine* 370.16 (2014), pp. 1514–1523. DOI: 10.1056/NEJMoa1310799. URL: https://www.ncbi.nlm.nih.gov/pubmed/24738668.

[103] J. Grützner et al. "Smart Diabetic Insole - Towards home feet health monitoring in order to prevent diabetic foot ulcer". In: *Biomedical Engineering/ Biomedizinische Technik*. Vol. 60. 1. 2015, pp. 252–253. URL: https://www.degruyter.com/downloadpdf/journals/bmte/60/s1/article-p1.pdf.

[104] Alessandra Guglielmi et al. "Statistical Methods in Medicine: Application to the Study of Glaucoma Progression". In: *Ocular Fluid Dynamics*. 2019, pp. 599–612. DOI: 10.1007/978-3-030-25886-3_24.

[105] Riccardo Guidotti et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5 (2018), pp. 1–42. DOI: 10.1145/3236009.

[106] Weina Guo, Mingyue Li, Yalan Dong, et al. "Diabetes is a risk factor for the progression and prognosis of COVID-19". In: *Diabetes/Metabolism Research and Reviews* 36.7 (2020), pp. 1–9. DOI: 10.1002/dmrr.3319.

[107] R. Gutekunst et al. "Ultrasonic diagnosis of the thyroid gland". In: *Deutsche Medizinische Wochenschrift* 113.27 (1988), pp. 1109–1112. DOI: 10.1055/s-2008-1067777.

[108] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3.Mar (2003). URL: https://www.jmlr.org/papers/v3/guyon03a.html, pp. 1157–1182.

[109] Jonathan B. S. Halford and Stewart D. Anderson. "Anxiety and depression in tinnitus sufferers". In: *Journal of Psychosomatic Research* 35.4/5 (1991), pp. 383–390. DOI: 10.1016/0022-3999(91)90033-K.

[110] Mark Hall. *HotSpot (Weka Package)*. URL: https : / / weka . sourceforge . io / packageMetaData / hotSpot / 1 . 0 . 4 . html. 2012. URL: https : / / weka . sourceforge . io / packageMetaData/hotSpot/1.0.4.html.

[111] Mark Hall et al. "The WEKA Data Mining Software: An Update". In: *ACM SIGKDD Explorations* 11.1 (2009), pp. 10–18. DOI: 10.1145/1656274.1656278.

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

[113] Jiawei Han, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". In: *ACM SIGMOD Record*. Vol. 29. 2. 2000, pp. 1–12. DOI: 10.1145/335191.335372.

[114] Tae Sun Han et al. "Gender Differences Affecting Psychiatric Distress and Tinnitus Severity". In: *Clinical Psychopharmacology and Neuroscience* 17.1 (2019), pp. 113–120. DOI: 10.9758/cpn.2019.17.1.113.

[115] Anke Hannemann, Nele Friedrich, Kathleen Dittmann, et al. "Age-and sex-specific reference limits for creatinine, cystatin C and the estimated glomerular filtration rate". In: *Clinical Chemistry and Laboratory Medicine* 50.5 (2012), pp. 919–926. DOI: 10.1515/CCLM.2011.788.

[116] Eduardo Hariton and Joseph J. Locascio. "Randomised controlled trials – the gold standard for effectiveness research". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 125.13 (2018), p. 1716. DOI: 10.1111/1471-0528.15199. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/1471-0528.15199. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.15199.

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

[118] 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/.

[119] Martin Hautzinger and Maja Bailer. *ADS-Allgemeine Depressionsskala*. Beltz, 2003.

[120] K. Hechenbichler and K. Schliep. "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification". In: *SFB 386, Ludwig-Maximilians University, Munich*. 399. 2004, pp. 1–16. URL: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9.

[121] J. L. Henry and P. H. 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). URL: https://europepmc.org/article/med/10753339, pp. 9–20.

[122] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020. URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[123] Franciso Herrera et al. "An overview on subgroup discovery: foundations and applications". In: *Knowledge and Information Systems* 29.3 (2011), pp. 495–525. DOI: 10.1007/s10115-010-0356-2.

[124] Gerhard Hesse. "Evidence and evidence gaps in tinnitus therapy". In: *GMS Current Topics in Otorhinolaryngology - Head and Neck Surgery* 15 (2016), pp. 1–42. DOI: 10.3205/cto000131.

[125] 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. DOI: 10.1016/j.eswa.2018.07.003. URL: https://doi.org/10.1016/j.eswa.2018.07.003.

[126] Tommy Hielscher et al. "Identifying Relevant Features for a Multi-factorial Disorder with Constraint-Based Subspace Clustering". In: *Computer-Based Medical Systems (CBMS)*. 2016, pp. 207–212. DOI: 10.1109/CBMS.2016.42.

[127] Tommy Hielscher et al. "Mining Longitudinal Epidemiological Data to Understand a Reversible Disorder". In: *Intelligent Data Analysis (IDA)*. 2014, pp. 120–130. DOI: 10.1007/978-3-319-12571-8_11.

[128] Tommy Hielscher et al. "Using Participant Similarity for the Classification of Epidemiological Data on Hepatic Steatosis". In: *Computer-Based Medical Systems (CBMS)*. 2014, pp. 1–7. DOI: 10.1109/CBMS.2014.28.

[129] Wolfgang Hiller and Gerhard Goebel. "Factors Influencing Tinnitus Loudness and Annoyance". In: *Archives of Otolaryngology–Head & Neck Surgery* 132.12 (2006), pp. 1323–1330. DOI: doi10.1001/archotol.132.12.1323. URL: https://jamanetwork.com/journals/jamaotolaryngology/article-abstract/484595.

[130] Aroon D. Hingorani, Daniëlle A. van der Windt, Richard D. Riley, et al. "Prognosis research strategy (PROGRESS) 4: Stratified medicine research". In: *BMJ: British Medical Journal* 346 (2013), pp. 1–9. DOI: 10.1136/bmj.e5793.

[131] 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. DOI: 10.2307/2685478.

[132] David C. Hoaglin. "John W. Tukey and Data Analysis". In: *Statistical Science* 18.3 (2003), pp. 311–318. DOI: 10.1214/ss/1076102418.

[133] 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), pp. 1–22. DOI: 10.1002/14651858.CD006371.pub3.

[134] Arthur E. Hoerl and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67. DOI: 10.2307/1271436.

[135] Albert Hofman, Monique M. B. Breteler, Cornelia M. van Duijn, et al. "The Rotterdam Study: 2010 objectives and design update". In: *European Journal of Epidemiology* 24.9 (2009), pp. 553–572. DOI: 10.1007/s10654-009-9386-z.

[136] Rolf Holle et al. "KORA–a research platform for population based health research". In: *Das Gesundheitswesen* 67.S1 (2005), pp. 19–25. DOI: 10.1055/s-2005-858235.

[137] Hyokyoung G. Hong, David C. Christiani, and Yi Li. "Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine". In: *Precision Clinical Medicine* 2.2 (2019), pp. 90–99. DOI: 10.1093/pcmedi/pbz007.

[138] M. Hörhold, B. F. Klapp, and U. Schimmack. "Testungen der Invarianz und der Hierarchie eines mehrdimensionalen Stimmungsmodells auf der Basis von Zweipunkterhebungen an Patienten-und Studentenstichproben". In: *Zeitschrift für Medizinische Psychologie* 2.1 (1993), pp. 27–35.

[139] 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.

[140] 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.

[141] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. URL: https://otexts.com/fpp2/. OTexts, 2018. URL: https://otexts.com/fpp2/.

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

[143] WHO MONICA Project Principal Investigators et al. "The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration". In: *Journal of Clinical Epidemiology* 41.2 (1988), pp. 105–114. DOI: 10.1016/0895-4356(88)90084-4.

[144] Tsunenori Ishioka. "An expansion of X-means for automatically determining the optimal number of clusters". In: *Computational Intelligence*. Vol. 2. URL: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.597. 2005, pp. 91–95.

[145] 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: 10.1089/thy.2011.0279.

[146] Avais Jabbar et al. "Thyroid hormones and cardiovascular disease". In: *Nature Reviews Cardiology* 14.1 (2017), pp. 39–55. DOI: 10.1038/nrcardio.2016.174.

[147] Isabelle Jaussent et al. "Insomnia Symptoms in Older Adults: Associated Factors and Gender Differences". In: *The American Journal of Geriatric Psychiatry* 19.1 (2011), pp. 88–97. DOI: 10.1097/JGP.0b013e3181e049b6.

[148] Zhenghui Gordon Jiang, Simon C. Robson, and Zemin Yao. "Lipoprotein metabolism in nonalcoholic fatty liver disease". In: *Journal of Biomedical Research* 27.1 (2013), pp. 1–13. DOI: 10.7555/JBR.27.20120077.

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

[150] L. Kaufman and P. J. Rousseeuw. "Clustering by Means of Medoids". In: *Statistical Data Analysis based on the L1-Norm and Related Methods* (1987), pp. 405–416.

[151] Jalil Kazemitabar et al. "Variable Importance Using Decision Trees". In: *Neural Information Processing Systems (NIPS)*. Vol. 30. 2017, pp. 426–435. URL: https://proceedings.neurips.cc/paper/2017/file/5737c6ec2e0716f3d8a7a5c4e0de0d9a-Paper.pdf.

[152] 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. DOI: 10.1016/j.amjcard.2012.08.012.

[153] 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. DOI: 10.1038/mp.2015.198.

[154] Kenji Kira and Larry A. Rendell. "The Feature Selection Problem: Traditional Methods and a New Algorithm". In: *AAAI Artificial Intelligence*. Vol. 2. URL: https://www.aaai.org/Library/AAAI/1992/aaai92-020.php. 1992, pp. 129–134. URL: https://www.aaai.org/Library/AAAI/1992/aaai92-020.php.

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

[156] Paul Klemm et al. "Interactive Visual Analysis of Image-Centric Cohort Study Data". In: *Transactions on Visualization and Computer Graphics (TVCG)* 20.12 (2014), pp. 1673–1682. DOI: 10.1109/TVCG.2014.2346591.

[157] Willi Klösgen and Michael May. "Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database". In: *Principles of Data Mining and Knowledge Discovery*. 2002, pp. 275–286. DOI: 10.1007/3-540-45681-3_23.

[158] Teuvo Kohonen. *Self-Organizing Maps*. Springer Science & Business Media, 2012. DOI: 10.1007/978-3-642-56927-2.

[159] Josua Krause, Adam Perer, and Kenney Ng. "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models". In: *Conference on Human Factors in Computing Systems (CHI)*. 2016, pp. 5686–5697. DOI: 10.1145/2858036.2858529.

[160] 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.

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

[162] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013. DOI: 10.1007/978-1-4614-6849-3.

[163] 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.

[164] Jens-Peter Kühn, Matthias Evert, Nele Friedrich, 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: 10.1097/RLI.0b013e31822b124c.

[165] M. J. Kumari, J. Subash, and S. Jagdish. "How to Prevent Amputation in Diabetic Patients". In: *International Journal of Nursing Education.* 6 (2014), pp. 40–44. DOI: https://doi.org/10.5958/0974-9357.2014.00602.3.

[166] 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), pp. 1–7. DOI: 10.1186/1472-6947-10-42.

[167] 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), pp. 1–46. DOI: 10.3389/fneur.2017.00046.

[168] Berthold Langguth et al. "Tinnitus and depression". In: *The World Journal of Biological Psychiatry* 12.7 (2011), pp. 489–500. DOI: 10.3109/15622975.2011.575178.

[169] Berthold Langguth et al. "Tinnitus: causes and clinical management". In: *The Lancet Neurology* 12.9 (2013), pp. 920–930. DOI: 10.1016/S1474-4422(13)70160-1.

[170] Cicero Augusto de Lara Pahins, Nivan Ferreira, and Joao Comba. "Real-Time Exploration of Large Spatiotemporal Datasets based on Order Statistics". In: *Transactions on Visualization and Computer Graphics (TVCG)* 26.11 (2019), pp. 3314–3326. DOI: 10.1109/TVCG.2019.2914446.

[171] 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: 10.1097/HJH.0b013e32833c211b.

[172] 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. DOI: 10.1227/NEU.0b013e3182503bf9.

[173] Lawrence A. Lavery et al. "Predictive value of foot pressure assessment as part of a population-based diabetes disease management program". In: *Diabetes Care* 26.4 (2003), pp. 1069–1073. DOI: 10.2337/diacare.26.4.1069.

[174] Nada Lavrač, Peter Flach, and Blaz Zupan. "Rule Evaluation Measures: A Unifying View". In: *Inductive Logic Programming (ILP)*. 1999, pp. 174–185. DOI: 10.1007/3-540-48751-4_17.

[175] Nada Lavrač et al. "Subgroup Discovery with CN2-SD". In: *Journal of Machine Learning Research* 5 (2004). URL: https://www.jmlr.org/papers/v5/lavrac04a.html, pp. 153–188.

[176] Matthijs van Leeuwen and Arno Knobbe. "Diverse subgroup set discovery". In: *Data Mining and Knowledge Discovery* 25.2 (2012), pp. 208–242. DOI: 10.1007/s10618-012-0273-y.

[177] Joffrey L. Leevy et al. "A survey on addressing high-class imbalance in big data". In: *Journal of Big Data* 5 (2018), pp. 1–42. DOI: 10.1186/s40537-018-0151-6.

[178] Pirmin Lemberger and Ivan Panico. *A Primer on Domain Adaptation*. 2020. eprint: https://arxiv.org/abs/2001.09994.

[179] 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. DOI: 10.1002/asmb.446.

[180] 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. DOI: 10.1007/s10620-006-9262-6.

[181] Hanna M. van Loo et al. "Major Depressive Disorder Subtypes to Predict Long-term Course". In: *Depression & Anxiety* 31.9 (2014), pp. 765–777. DOI: 10.1002/da.22233.

[182] Alessandra Lugo et al. "Sex-Specific Association of Tinnitus With Suicide Attempts". In: *JAMA Otolaryngology–Head & Neck Surgery* 145.7 (2019), pp. 685–687. DOI: 10.1001/jamaoto.2019.0566. URL: https://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2732497#old190004r3.

[183] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Neural Information Processing Systems (NIPS)*. URL: https://dl.acm.org/doi/10.5555/3295222.3295230. 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[184] Scott M. Lundberg et al. *Explainable AI for Trees: From Local Explanations to Global Understanding*. 2019. eprint: https://arxiv.org/abs/1905.04610.

[185] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008). URL: https://www.jmlr.org/papers/v9/vandermaaten08a.html, pp. 2579–2605.

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

[187] 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: 10.1161/ATVBAHA.112.300556.

[188] Pablo Martinez-Devesa et al. "Cognitive behavioural therapy for tinnitus". In: *Cochrane Database of Systematic Reviews* 9 (2010). DOI: 10.1002/14651858.CD005233.pub3. URL: https://doi.org//10.1002/14651858.CD005233.pub3.

[189] M. Pilar Matud. "Gender differences in stress and coping styles". In: *Personality and Individual Differences* 37.7 (2004), pp. 1401–1415. DOI: 10.1016/j.paid.2004.01.010.

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

[191] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform manifold approximation and projection for dimension reduction*. 2018. eprint: https://arxiv.org/abs/1802.03426.

[192] Carmen P. McLean et al. "Gender Differences in Anxiety Disorders: Prevalence, Course of Illness, Comorbidity and Burden of Illness". In: *Journal of Psychiatric Research* 45.8 (2011), pp. 1027–1035. DOI: 10.1016/j.jpsychires.2011.03.006.

[193] Corine Meric et al. "Psychopathological profile of tinnitus sufferers: evidence concerning the relationship between tinnitus features and impact on life". In: *Audiology and Neurotology* 3.4 (1998), pp. 240–252. DOI: 10.1159/000013796.

[194] David Meyer et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien*. R package version 1.7-3. 2019. URL: https://CRAN.R-project.org/package=e1071.

[195] Antao Ming et al. "Study protocol for a randomized controlled trial to test for preventive effects of diabetic foot ulceration by telemedicine that includes sensor-equipped insoles combined with photo documentation". In: *Trials* 20.1 (2019), pp. 1–12. DOI: 10.1186/s13063-019-3623-x.

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

[197] Probal K. Moulik, Robert Mtonga, and Geoffrey V. Gill. "Amputation and mortality in new-onset diabetic foot ulcers stratified by etiology". In: *Diabetes Care* 26.2 (2003), pp. 491–494. DOI: 10.2337/diacare.26.2.491.

[198] 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), pp. 1–13. DOI: 10.1038/srep11104.

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

[200] Uli Niemann et al. "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. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0161326.

[201] Uli Niemann et al. "Phenotyping chronic tinnitus patients using self-report questionnaire data: cluster analysis and visual comparison". In: *Scientific Reports* 10.1 (2020), pp. 1–10. DOI: 10.1038/s41598-020-73402-8. URL: https://doi.org/10.1038/s41598-020-73402-8.

[202] Uli Niemann et al. "Plantar temperatures in stance position: A comparative study with healthy volunteers and diabetes patients diagnosed with sensoric neuropathy". In: *EBioMedicine* 54.102712 (2020), pp. 1–11. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2020.102712. URL: http://www.sciencedirect.com/science/article/pii/S2352396420300876.

[203] Uli Niemann et al. "Rupture Status Classification of Intracranial Aneurysms Using Morphological Parameters". In: *Computer-Based Medical Systems (CBMS)*. 2018, pp. 48–53. DOI: 10.1109/CBMS.2018.00016.

[204] Uli Niemann et al. "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. URL: https://doi.org/10.1371/journal.pone.0228037.

[205] Susan Nolen-Hoeksema. "Gender Differences in Depression". In: *Current Directions in Psychological Science* 10.5 (2001), pp. 173–176. DOI: 10.1111/1467-8721.00142.

[206] Dominic Oliver et al. "What Causes the Onset of Psychosis in Individuals at Clinical High Risk? A Meta-analysis of Risk and Protective Factors". In: *Schizophrenia Bulletin* 46.1 (2020), pp. 110–120. DOI: 10.1093/schbul/sbz039.

[207] Ahmed Oussous et al. "Big Data technologies: A survey". In: *Journal of King Saud University-Computer and Information Sciences* 30.4 (2018), pp. 431–448. DOI: 10.1016/j.jksuci.2017.06.001.

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

[209] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[210] Mykola Pechenizkiy et al. "Heart Failure Hospitalization Prediction in Remote Patient Management Systems". In: *Computer-Based Medical Systems (CBMS)*. 2010, pp. 44–49. DOI: 10.1109/CBMS.2010.6042612.

[211] Dan Pelleg and Andrew W. Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". In: *International Conference on Machine Learning (ICML)*. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.3377. 2000, pp. 727–734.

[212] John S. Phillips and Don McFerran. "Tinnitus retraining therapy (TRT) for tinnitus". In: *Cochrane Database of Systematic Reviews* 3 (2010), pp. 1–16. DOI: 10.1002/14651858.CD007330.pub2.

[213] Marco Piccinelli and Greg Wilkinson. "Gender differences in depression: Critical review". In: *The British Journal of Psychiatry* 177.6 (2000), pp. 486–492. DOI: 10.1192/bjp.177.6.486.

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

[215] Patricia Ciminelli Linhares Pinto, Tanit Ganz Sanchez, and Shiro Tomita. "The impact of gender, age and hearing loss on tinnitus severity". In: *Brazilian Journal of Otorhinolaryngology* 76.1 (2010), pp. 18–24. DOI: 10.1590/S1808-86942010000100004. URL: https://www.sciencedirect.com/science/article/pii/S1808869415313483.

[216] Thierry Poynard et al. "Apolipoprotein AI and alcoholic liver disease". In: *Hepatology* 6.6 (1986), pp. 1391–1395. DOI: 10.1002/hep.1840060628.

[217] Bernhard Preim and Kai Lawonn. "A Survey of Visual Analytics for Public Health". In: *Computer Graphics Forum*. Vol. 39. 1. 2020, pp. 543–580. DOI: 10.1111/cgf.13891.

[218] Bernhard Preim et al. "Visual Analytics for Epidemiological Cohort Studies". In: *Eurographics Medical Price*. 2019. URL: http://www.vismd.de/lib/exe/fetch.php?media=files:misc:preim_2019_medp.pdf.

[219] Bernhard Preim et al. "Visual Analytics of Image-Centric Cohort Studies in Epidemiology". In: *Visualization in Medicine and Life Sciences III*. Springer International Publishing, 2016, pp. 221–248. DOI: 10.1007/978-3-319-24523-2_10.

[220] John R. Quinlan. "Learning with Continuous Classes". In: *Artificial Intelligence (AI)*. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.885. 1992, pp. 343–348.

[221] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[222] 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. DOI: 10.1177/014662167700100306.

[223] Dheeraj Raju et al. "Exploring factors associated with pressure ulcers: A data mining approach". In: *International Journal of Nursing Studies* 52.1 (2014), pp. 102–111. DOI: 10.1016/j.ijnurstu.2014.08.002. URL: http://www.sciencedirect.com/science/article/pii/S0020748914002053.

[224] G. Rayman et al. "A study of factors governing fluid filtration in the diabetic foot". In: *European Journal of Clinical Investigation* 24.12 (1994), pp. 830–836. ISSN: 0014-2972 (Print) 0014-2972 (Linking). DOI: 10.1111/j.1365-2362.1994.tb02027.x. URL: https://www.ncbi.nlm.nih.gov/pubmed/7705378%20http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2362.1994.tb02027.x/abstract.

[225] G. Rayman et al. "Microvascular response to tissue injury and capillary ultrastructure in the foot skin of type I diabetic patients". In: *Clinical Science* 89.5 (1995), pp. 467–474. ISSN: 0143-5221 (Print) 0143-5221 (Linking). DOI: 10.1042/cs0890467. URL: https://www.ncbi.nlm.nih.gov/pubmed/8549060%20http://www.clinsci.org/content/89/5/467.long.

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

[227] 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. DOI: 10.2105/AJPH.2004.047225.

[228] Loredana Rizzo et al. "Custom-made orthesis and shoes in a structured follow-up program reduces the incidence of neuropathic ulcers in high-risk diabetic foot patients". In: *International Journal of Lower Extremity Wounds* 11.1 (2012), pp. 59–64. DOI: 10.1177/1534734612438729.

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

[230] Sylvia Saalfeld et al. "Semiautomatic neck curve reconstruction for intracranial aneurysm rupture risk assessment based on morphological parameters". In: *International Journal of Computer Assisted Radiology and Surgery (IJCARS)* 13.11 (2018), pp. 1781–1793. DOI: 10.1007/s11548-018-1848-x.

[231] James W. Salazar et al. "Depression in Patients with Tinnitus: A Systematic Review". In: *Otolaryngology–Head and Neck Surgery* 161.1 (2019), pp. 28–35. DOI: 10.1177/0194599819835178.

[232] Wojciech Samek et al. *Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond*. 2020. eprint: https://arxiv.org/abs/2003.07631.

[233] Jon Sánchez-Valle, Hector Tejero, José María Fernández, et al. "Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships". In: *Nature Communications* 11.1 (2020), pp. 1–13. DOI: 10.1038/s41467-020-16540-x.

[234] Peter Sarlin. "Self-organizing time map: An abstraction of temporal multivariate patterns". In: *Neurocomputing* 99.1 (2013), pp. 496–508. DOI: 10.1016/j.neucom.2012.07.011.

[235] N. C. Schaper et al. "Prevention and management of foot problems in diabetes: A Summary Guidance for Daily Practice 2015, based on the IWGDF guidance documents". In: *Diabetes Research and Clinical Practice* 124 (2017), pp. 84–92. ISSN: 1872-8227 (Electronic) 0168-8227 (Linking). DOI: 10.1016/j.diabres.2016.12.007. URL: https://www.ncbi.nlm.nih.gov/pubmed/28119194.

[236] 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.

[237] Winfried Schlee et al. "Visualization of Global Disease Burden for the Optimization of Patient Management and Treatment". In: *Frontiers in Medicine* 4 (2017), pp. 1–12. DOI: 10.3389/fmed.2017.00086.

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

[239] Bernhard Schölkopf. *Causality for Machine Learning*. 2019. eprint: https://arxiv.org/abs/1911.10500.

[240] Gudrun Scholler, Herbert Fliege, and Burghard F. Klapp. "Fragebogen zu Selbstwirksamkeit, Optimismus und Pessimismus". In: *Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)* 49.8 (1999), pp. 275–283. DOI: 10.23668/psycharchives.337.

[241] Nicholas J. Schork. "Personalized medicine: Time for one-person trials". In: *Nature News* 520.7549 (2015), pp. 609–611. DOI: 10.1038/520609a.

[242] 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.

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

[244] Claudia Seydel et al. "Gender and Chronic Tinnitus: Differences in Tinnitus-Related Distress Depend on Age and Duration of Tinnitus". In: *Ear and Hearing* 34.5 (2013), pp. 661–672. DOI: 10.1097/AUD.0b013e31828149f2.

[245] Lloyd S. Shapley. "A Value for n-Person Games". In: *Contributions to the Theory of Games* 2.28 (1953), pp. 307–317. DOI: 10.1515/9781400881970-018.

[246] Smadar Shilo, Hagai Rossman, and Eran Segal. "Axes of a revolution: challenges and promises of big data in healthcare". In: *Nature Medicine* 26 (2020), pp. 29–38. DOI: 10.1038/s41591-019-0727-5.

[247] Chaitanya Shivade et al. "A review of approaches to identifying patient phenotype cohorts using electronic health records". In: *Journal of the American Medical Informatics Association* 21.2 (2014), pp. 221–230. DOI: 10.1136/amiajnl-2013-001935.

[248] Zaigham Faraz Siddiqui et al. "Predicting the post-treatment recovery of the patients suffering from TBI". In: *Brain Informatics* 2 (2015), pp. 33–44. DOI: 10.1007/s40708-015-0010-6.

[249] Judith D. Singer and John B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, 2003. DOI: 10.1093/acprof:oso/9780195152968.001.0001.

[250] Nalini Singh, David G. Armstrong, and Benjamin A. Lipsky. "Preventing foot ulcers in patients with diabetes". In: *Journal of the American Medical Association* 293.2 (2005), pp. 217–228. DOI: 10.1001/jama.293.2.217.

[251] Shalini S. Singh and N. C. Chauhan. "K-means v/s K-medoids: A Comparative Study". In: *National Conference on Recent Trends in Engineering & Technology*. 2011, pp. 839–844.

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

[253] William W. Stead. "Clinical Implications and Challenges of Artificial Intelligence and Deep Learning". In: *Journal of the American Medical Association* 320.11 (2018), pp. 1107–1108. DOI: 10.1001/jama.2018.11029.

[254] Felix Stickel, Stephan Buch, Katharina Lau, 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: 10.1002/hep.24017.

[255] 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. DOI: 10.1007/s10115-013-0679-x.

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

[257] P. C. Sun et al. "Impaired microvascular flow motion in subclinical diabetic feet with sudomotor dysfunction". In: *Microvascular Research* 83.2 (2012), pp. 243–248. ISSN: 1095-9319 (Electronic) 0026-2862 (Linking). DOI: 10.1016/j.mvr.2011.06.002. URL: https://www.ncbi.nlm.nih.gov/pubmed/21722653%20https://ac.els-cdn.com/S0026286211001038/1-s2.0-S0026286211001038-main.pdf?_tid=cb67baf2-0f5d-11e8-ab5f-00000aab0f27&acdnat=1518375795_1830bdee5053817ea502a7930bec13ca.

[258] Noboru Takamura et al. "Thyroid function is associated with carotid intima-media thickness in euthyroid subjects". In: *Atherosclerosis* 204.2 (2009), pp. 77–81. DOI: 10.1016/j.atherosclerosis.2008.09.022.

[259] Pang-Ning Tan et al. *Introduction to Data Mining*. Second edition. Pearson, 2019.

[260] 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: 10.1056/NEJMra0912063.

[261] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2020. URL: https://www.R-project.org/.

[262] Sarah M. Theodoroff et al. "Individual Patient Factors Associated with Effective Tinnitus Treatment". In: *Journal of the American Academy of Audiology* 25.7 (2014), pp. 631–643. DOI: 10.3766/jaaa.25.7.2.

[263] 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.

[264] Matthew S. Thiese. "Observational and interventional study design types; an overview". In: *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.

[265] Sana Tonekaboni et al. "What Clinicians Want: Contextualizing Explainable Machine-Learning for Clinical End Use". In: *Machine Learning for Healthcare*. URL: http://proceedings.mlr.press/v106/tonekaboni19a.html. 2019, pp. 359–380. URL: http://proceedings.mlr.press/v106/tonekaboni19a.html.

[266] 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. DOI: 10.1016/j.cpr.2017.12.006.

[267] 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.

[268] Anastasios A. Tsiatis. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press, 2019.

[269] Richard Tyler et al. "Identifying Tinnitus Subgroups With Cluster Analysis". In: *American Journal of Audiology* 17.2 (2 2008), pp. 176–184. DOI: 10.1044/1059-0889(2008/07-0044).

[270] 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), pp. 144–149. DOI: 10.4274/jcrpe.3719.

[271] Vishnu Unnikrishnan et al. "Entity-level stream classification: exploiting entity similarity to label the future observations referring to an entity". In: *International Journal of Data Science and Analytics* 9.1 (2020), pp. 1–15. DOI: 10.1007/s41060-019-00177-1.

[272] 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. DOI: 10.1007/978-3-642-41016-1_27.

[273] Annemarie Van der Wal et al. "Sex Differences in the Response to Different Tinnitus Treatment". In: *Frontiers in Neuroscience* 14.422 (2020), pp. 1–9. DOI: 10.3389/fnins.2020.00422. URL: https://www.frontiersin.org/article/10.3389/fnins.2020.00422.

[274] Alfredo Vellido. "The importance of interpretability and visualization in Machine Learning for applications in medicine and health care". In: *Neural Computing and Applications* 32 (2019), pp. 18069–18083. DOI: 10.1007/s00521-019-04051-w.

[275] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth edition. Springer, 2002. DOI: 10.1007/978-0-387-21706-2. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[276] A. Veves et al. "The risk of foot ulceration in diabetic patients with high foot pressure: a prospective study". In: *Diabetologia* 35.7 (1992), pp. 660–663. DOI: 10.1007/BF00400259.

[277] Marco Viceconti, Peter Hunter, and Rod Hose. "Big Data, Big Knowledge: Big Data for Personalized Healthcare". In: *Biomedical and Health Informatics* 19.4 (2015), pp. 1209–1215. DOI: 10.1109/JBHI.2015.2406883.

[278] 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. DOI: 10.1044/2014_JSLHR-S-12-0193.

[279] Giorgio Visani, Enrico Bagli, and Federico Chesani. *OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms*. 2020. eprint: https://arxiv.org/abs/2006.05714.

[280] M. Volmer-Thole and R. Lobmann. "Neuropathy and Diabetic Foot Syndrome". In: *International Journal of Molecular Sciences* 17.6 (2016), pp. 1–11. ISSN: 1422-0067 (Electronic) 1422-0067 (Linking). DOI: 10.3390/ijms17060917. URL: https://www.ncbi.nlm.nih.gov/pubmed/27294922.

[281] Henry Völzke. "Multicausality in fatty liver disease: is there a rationale to distinguish between alcoholic and non-alcoholic origin?" In: *World Journal of Gastroenterology* 18.27 (2012), pp. 3492–3501. DOI: 10.3748/wjg.v18.i27.3492.

[282] Henry Völzke, Dietrich Alte, Carsten Oliver Schmidt, et al. "Cohort Profile: The Study of Health in Pomerania". In: *International Journal of Epidemiology* 40.2 (2011), pp. 294–307. DOI: 10.1093/ije/dyp394.

[283] Henry Völzke, Glenn Fung, Till Ittermann, et al. "A new, accurate predictive model for incident hypertension". In: *Journal of Hypertension* 31.11 (2013), pp. 2142–2150. DOI: 10.1097/HJH.0b013e328364a16d.

[284] Henry Völzke et al. "Hepatic steatosis is associated with an increased risk of carotid atherosclerosis". In: *World Journal of Gastroenterology* 11.12 (2005), pp. 1848–1853. DOI: 10.3748/wjg.v11.i12.1848.

[285] Henry Völzke et al. "Menopausal status and hepatic steatosis in a general female population". In: *Gut* 56.4 (2007), pp. 594–595. DOI: 10.1136/gut.2006.115345.

[286] 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), pp. 185–192. DOI: 10.3238/arztebl.2015.0185.

[287] Roelof Waaijman et al. "Risk Factors for Plantar Foot Ulcer Recurrence in Neuropathic Diabetic Patients". In: *Diabetes Care* 37.6 (2014), pp. 1697–1705. DOI: 10.2337/dc13-2470.

[288] Ute Waldecker. "Pedographic classification and ulcer detection in the diabetic foot". In: *Foot and Ankle Surgery* 18.1 (2012), pp. 42–49. DOI: 10.1016/j.fas.2011.03.004.

[289] Ron Wehrens and Johannes Kruisselbrink. "Flexible Self-Organizing Maps in kohonen 3.0". In: *Journal of Statistical Software* 87.1 (2018), pp. 1–18. DOI: 10.18637/jss.v087.i07.

[290] 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.

[291] Marieke J. H. 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. DOI: 10.1161/01.STR.0000260955.51401.cd.

[292] 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. DOI: 10.1046/j.1525-1497.1997.00076.x.

[293] G. Wiesner and E. K. 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. DOI: 10.1007/s00103-003-0793-0.

[294] D. Randall Wilson and Tony R. Martinez. "Improved Heterogeneous Distance Functions". In: *Journal of Artificial Intelligence Research* 6.1 (1997), pp. 1–34. DOI: 10.1613/jair.346.

[295] Krist Wongsuphasawat and David Gotz. "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization". In: *Transactions on Visualization and Computer Graphics (TVCG)* 18.12 (2012), pp. 2659–2668. DOI: 10.1109/TVCG.2012.225.

[296] 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.

[297] Jianping Xiang et al. "Hemodynamic-Morphologic Discriminants for Intracranial Aneurysm Rupture". In: *Stroke* 42.1 (2011), pp. 144–152. DOI: 10.1161/STROKEAHA.110.592923.

[298] Guo Yu, Daniela Witten, and Jacob Bien. *Controlling Costs: Feature Selection on a Budget*. 2020. eprint: https://arxiv.org/abs/1910.03627.

[299] Kui Yu et al. "Causality-based Feature Selection: Methods and Evaluations". In: *ACM Computing Surveys* 53.5 (2020), pp. 1–36. DOI: 10.1145/3409382. URL: https://dl.acm.org/doi/pdf/10.1145/3409382.

[300] Xin Yuan et al. "Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes". In: *The American Journal of Human Genetics* 83.4 (2008), pp. 520–528. DOI: 10.1016/j.ajhg.2008.09.012.

[301] Zhiyuan Zhang, David Gotz, and Adam Perer. "Iterative cohort analysis and exploration". In: *Information Visualization* 14.4 (2015), pp. 289–307. DOI: 10.1177/1473871614526077.

[302] 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: 10.1016/j.artmed.2013.04.008.

[303] Juan Zhao et al. "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction". In: *Scientific Reports* 9.1 (2019), pp. 1–10. DOI: 10.1038/s41598-018-36745-x.