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

Uli Niemann

23.06.2020

Contents

Structure 4

1 Thesis overview **7**

1.1 Motivation 7

1.2 Medical applications 8

1.3 Open challenges 9

1.4 Structure and summary of contributions 10

I SUBPOPULATION DISCOVERY IN HIGH-DIMENSIONAL DATA **11**

2 Interactive discovery and inspection of subpopulations **13**

2.1 Motivation and comparison with related work 13

2.2 The SHIP dataset 14

2.3 Subpopulation discovery workflow and interactive mining assistant 14

2.4 Experiments and findings 14

2.5 Benefits of our workflow 14

3 Identifying diverse subpopulations **15**

3.1 Motivation and comparison with related work 15

3.2 Finding diverse classification rules 15

3.3 Experiments and findings 15

3.4 Benefits our our method 15

II EXPLOITING TEMPORAL INFORMATION **17**

4 Constructing evolution features to capture change over time **19**

4.1 Motivation and comparison with related work 19

4.2 Overview of the mining workflow 20

4.3 Generating evolution features 20

4.4 Experiments and findings 20

4.5 Benefits our our method 20

5 Feature extraction from short temporal sequences for clustering 21

5.1 Motivation and comparison with related work . . . 21
5.2 The diabatic foot dataset . . . 21
5.3 Overview of the mining workflow . . . 21
5.4 Experiments and findings . . . 21
5.5 Benefits our our method . . . 21

III POST-MINING FOR INTERPRETATION 23

6 Post-hoc interpretation of classification models 25

6.1 Motivation and comparison with related work . . . 25
6.2 The aneurysm dataset . . . 26
6.3 The tinnitus dataset . . . 26
6.4 Overview of the mining workflow . . . 26
6.5 Experiments and findings on aneurysm data . . . 26
6.6 Experiments and findings on tinnitus data . . . 26
6.7 Benefits our our method . . . 26

IV SUMMARY 27

7 Conclusion and future work 29

7.1 Research results for medical expert-guided knowledge discovery . . . 29
7.2 Future work . . . 29

V APPENDIX 31

Structure

1 Thesis Overview
1.1 Motivation
1.2 Medical Applications
1.3 Open Challenges
1.4 Structure and Summary of Contributions

PART I SUBPOPULATION DISCOVERY

2 Interactive discovery and inspection of subpopulations
2.1 The SHIP dataset
2.2 Motivation and comparison with related work
2.3 Subpopulation discovery workflow and interactive mining assistant
2.4 Experiments and findings

2.5 Benefits of our workflow

3 Identifying diverse subpopulations
3.1 Motivation and comparison with related work
3.2 Finding diverse classification rules
3.3 Experiments and findings
3.4 Benefits our our method

PART II EXPLOITING TEMPORAL INFORMATION

4 Constructing evolution features to capture change over time
4.1 Motivation and comparison with related work
4.2 Overview of the mining workflow
4.3 Generating evolution features
4.4 Experiments and findings
4.5 Benefits our our method

5 Feature extraction from short temporal sequences for clustering
5.1 The diabatic foot dataset
5.1 Motivation and comparison with related work 5.2 Overview of the mining workflow
5.3 Experiments and findings
5.4 Benefits our our method

PART III POST-MINING FOR INTERPRETATION

6 Post-hoc interpretation of classification models
6.1 Motivation and comparison with related work
6.2 The aneurysm dataset
6.3 The tinnitus dataset
6.2 Overview of the mining workflow
6.3 Experiments and findings on aneurysm data
6.4 Experiments and findings on tinnitus data
6.5 Benefits our our method

PART IV SUMMARY

7 Conclusion and future work
7.1 Research results for medical expert-guided knowledge discovery
7.2 Future work

PART V APPENDIX

Bibliography
List of publications
Abbreviations

Chapter 1

Thesis overview

1.1 Motivation

The analysis of medical data typically strives towards (1) the identification of long-term determinants and protective factors for an outcome of interest, (2) the discovery of subpopulations with increased disease prevalence, and (3) the generation of robust models capable of explaining relationships between one or several independent variables and the outcome. For example, epidemiologists aim to discover causal associations between multiple risk factors and an outcome in cohort studies collecting data which encompass rich information about the participants, gathered from questionnaires, medical examinations, laboratory analyses and image acquisition. Often, these data are repeatedly collected over time, for example in longitudinal studies. Thus, researchers incorporate a participants history in their analyses to account for temporal aspects.
Finding causal associations between variables commonly requires medical scholars to first carefully derive some hypotheses from clinical daily routine, experimental studies or extensive literature research, and then formally test them on statistical significance. However, with the ever-growing volume and heterogeneity of medical data, traditional hypothesis-driven workflows become increasingly impractical because some important inherent associations between variables might remain undiscovered.

Data Mining (DM) and Machine Learning[1] (ML) can enhance medical research by discovering subpopulations of study participants which are similar with respect to the outcome, and thus can be used to generate new hypotheses. The proliferation of medical ML applications is triggered by multiple reasons, such as the desire of exploiting the plethora of gathered study participant information and the ubiquitousness of "A.I." success stories in the media. However, simply building complex data-driven models does not guarantee that knowledge can be derived without effort. Most state-of-the-art ML algorithms such as deep neuronal networks and gradient boosting machines generate so called black-box models with multiple layers of complexity incorporating many multi-variate, non-linear interactions among variables, which are hard to present intuitively. It is vital that the application expert, which is not a practitioner, but a scholar working in a clinical or epidemiological environment, is equipped with means to understand, explore and visualize the models, so that s/he can drill down to specific individual patterns and gain actionable insights that ultimately contribute to prevention, diagnosis and treatment of diseases. Since health care data can be derived from diverse sources such as epidemiological studies or clinical trials, major characteristics of collected datasets vary which requires an adaption of methods tailored to the characteristics of the application scenario.

The goal of this thesis is to support medical experts by providing an intelligent assistance for the analysis of high-dimensional time-stamped medical data. During the phases of model generation and model post-processing, several challenges have to be dealt with, so that the expert is able to derive actionable knowledge. These challenges translate to the following four demands:

[1] The terms Data Mining and Machine Learning are used synonymously.

(1) *Interpretability of Patterns*: patterns, i.e. clusters, rules, classification models, must be interpretable; the process of pattern generation must be interpretable.
(2) *Exploitation of Time*: time must be exploited, while satisfying requirement (1). Medical scholars search for long-term determinants of severe diseases. Finding patterns in the patient history and evolution can contribute to this aim.
(3) *Minimization of Redundancy*: redundancy must be minimized, while satisfying requirement (1). Patterns might be overlapping with respect to the subjects they cover. This leads to a redundancy of patterns which negatively affects the perceived quality of the model. One task is to minimize redundancy and deliver only a list of relevant patterns to the expert.
(4) *Validation*: validation must be part of the analysis, while satisfying (1), (2) and (3).

There are several aspects to deal with:

1. Dimensionality: to what extend can the same methods be used on low-dimensional and high-dimensional data?
2. Time granularity: to what extend can the same methods be used on data with few discrete waves and with dense time-stamped recordings?
3. The impact of the protocol: both data recorded for specific research questions (DIAB, Dan-Monica, AneurD) are investigated as well as data recorded independently of the research questions (SHIP, TinD). To what extend can the same approaches be used, what deviations are needed?

1.2 Medical applications

A categorization of time-stamped medical data. The structure of medical data naturally depends on the study design. In this thesis proposal, medical data is categorized using the following five criteria.

- *Number of variables*: number of variables, features, attributes or dimensions
- *Number of individuals*: number of individuals, subjects, instances, patients or study participants
- *Temporal horizon*: timespan from first to last assessment
- *Number of assessments over time*: number of repeated assessments over time
- *Temporal granularity*: frequency of repeated assessments

Table 1.1 juxtaposes the five application datasets with respect to the five criteria depicted above.

Table 1.1: Study dataset characteristics.

	Number of variables	Number of individuals	Temporal horizon	Number of assessments	Temporal granularity
SHIP	↑ (>100)	→ (>800)	↑ (~15 yr)	↓ (3 obs.)	↓ (1 ass./5 yr)
DFSS	↓ (<10)	↓ (<60)	↓ (~100 min)	↑ (> 100 obs.)	↑ (≥ 3 rec./min)
AneurD	↓ (<30)	→ (100)	• (none)	• (none)	• (none)
Dan-MONICA	↑ (>100)	→ (>1000)	↑ (~10 yr)	↓ (3 obs.)	↓ (1 ass./5 yr)
TinD	↑ (>100)	→ (>1000)	↓ (~10 days)	↑ (> 2 obs.)	NA

We will evaluate our methods on five medical datasets.
SHIP. The "Study of Health in Pomerania" (SHIP) [1] encompasses two independent cohorts, SHIP-CORE and SHIP-TREND, consisting of participants aged from 20 to 79 years with main residency in the study region. Baseline examinations for SHIP-CORE were performed between 1997 and 2001 (SHIP-0, n=4308). Three follow-up examinations were conducted in intervals of 5 years, continuously adding new variables including MRI scans beginning from

the second follow-up (SHIP-2). Baseline information for the SHIP-TREND (TREND-0, n=4420) was collected in 2008-2012. In both cohorts, a broad range of participant information was gathered, including socio-demographics, lifestyle and consumption behavior, medication intake, anthropometric values, blood and urine laboratory values, whole-body magnet resonance imaging, and others.

DFSS. While SHIP investigates associations between risk factors of common diseases and health-related conditions in general, the "Diabetic Foot Syndrome' ' study [2], [3] is a case-control study specifically investigating differences between diabetes patients diagnosed with severe polyneuropathy and healthy controls. In a pre-specified sequence of actions alternating between two postures, standing and sitting, custom-made sensor-bearing insoles placed in a closed shoe measured changes of a subject's plantar pressures and temperatures with high frequency over time. These measurements allowed for the analysis of pathophysiological and sensation-related differences between the two groups.

Aneurysm database (AneurD). The "aneurysm database" contains angiographical image data of 74 patients with a total of 100 intracranial aneurysms [4]. Intracranial aneurysms are pathologic dilations of the intracranial vessel wall, bearing the risk of rupture and thus subarachnoidal hemorrhages with often fatal consequences for the patient. Since treatment may cause severe complications as well, the goal is to generate models that can separate between ruptured and non-ruptured aneurysms.

MONICA. This dataset is derived from the Danish population of the WHO project MONICA (Multinational MONItoring of trends and determinants in CArdiovascular disease). MONICA is a longitudinal epidemiological cohort study focusing on risk factors of cardio-vascular diseases, as well as causes and trends regarding the differences in the mortality between countries [5]. Overall, 37 centers from 21 countries participated in the cohort study conducted between 1976 and 2002. Accomplishments of the project include the establishment of a profound database for prevention research and improvements on treatment of cardio-vascular diseases due to the identification of risk factors such as increased cholesterol and nicotine levels. In the Danish MONICA cohort, over 30,000 residents from the south-western of Copenhagen County, aged between 25 and 74 years, participated in three waves from 1982 to 1991 [6]. The baseline examinations were conducted in 1982-1986 (DAN-MONICA I) with the first two follow-ups in 1986-1987 (DAN-MONICA II) and in 1991-1992 (DAN-MONICA III).

Tinnitus dataset (TinD). This dataset contains information on a cohort of a total of 3,971 tinnitus patients who had been treated at the tinnitus centre of the university medicine in Berlin between January 2011 and October 2015. All included patients had been suffering from tinnitus for 3 months or longer, were 18 years of age or older and had sufficient knowledge of the German language. At registration, patients were asked to complete 15 questionnaires on socio-demographics, tinnitus-related distress, frequency, loudness, localization and quality as well as physical and mental health status, including levels of depression and perceived stress. To assess efficacy of the treatment, patients filled the same battery of questionnaires a second time.

1.3 Open challenges

The motivation described in the previous section leads to the following research questions, including the core research question (RQ). The core research question of the thesis is: **How to generate accurate yet understandable patterns for subpopulation discovery in high-dimensional time-stamped medical data?** This question is central for the thesis. Since it includes multiple aspects, such as feature engineering, model learning, and post-mining steps, the question is subdivided into the following three research questions.

RQ 1 How to leverage understandable models for high-dimensional time-stamped medical data?

RQ 2 How to effectively capture and exploit implicit temporal information to improve model quality?

RQ 3 How to extract, process and display the most relevant (temporal) patterns for an efficient expert-driven model exploration?

Table 1.2 shows a mapping between scientific articles associated with this thesis and addressed research questions.

1.4 Structure and summary of contributions

Contributions of the published and submitted work include

- a workflow and interactive application for data preparation, mining, and inspection of patterns for epidemiological data [7], [8],
- a feature engineering framework for modeling and exploitation of the cohort participant evolution over time to improve classification quality [9], [10],
- a mechanism for feature extraction from raw multi-variate time-series to identify relevant subgroups in patients and healthy controls and [2], [11],
- a clustering-based algorithm that hierarchically reorganizes classification rule sets and summarizes all important concepts wihle maintaining diversity between the rule clusters to reduce pattern redundancy [12],
- a pipeline for feature extraction, classification with post-mining functionalities such as feature ranking and inspection [4], and
- results that confirm findings from medical literature or results that can be validated in independent studies.

Table 1.2: Mapping of published and planned articles to research questions.

Article	**RQ1**	**RQ2**	**RQ3**	Datasets
[7]	✓		✓	SHIP
[8]	✓		✓	SHIP
[13]	✓			SHIP
[9]	✓	✓		SHIP
[11]	✓	✓		DFSS
[2]	✓	✓		DFSS
[12]	✓			SHIP
[4]	✓			AneurD
[14]	✓	✓		SHIP, MONICA
[10]	✓		✓	TinD
[3]	✓		✓	DFSS

Part I

SUBPOPULATION DISCOVERY IN HIGH-DIMENSIONAL DATA

Chapter 2

Interactive discovery and inspection of subpopulations

Based on [7]

2.1 Motivation and comparison with related work

Medical decisions on diagnosis and treatment of multifactorial diseases are based on clinical and epidemiological studies. The latter accommodate information on participants with and without the disorder and allow for discriminative model learning and, in the longitudinal design, for understanding the progress of a disorder (possibly towards a disease). For example, there are several studies on the identification of factors (like obesity or alcohol consumption) and outcomes (like cardiovascular diseases) associated with hepatic steatosis. Findings on genetic and non-genetic factors include [15], [16], [17]; findings on associated outcomes include [18] and [19]. However, these studies identified risk factors and/or associated outcomes that pertain to the whole population. Our work emanates from the necessity to identify such factors and outcomes for subpopulations and thus to stimulate personalized diagnosis and treatment, as expected in personalized medicine [20], [21].

Classification on subpopulations was studied by Zhanga and Kodell in [22]. They pointed out that the complete population can be very heterogeneous, so that classifier performance on the whole dataset can be low. Therefore, they first trained an ensemble of classifiers, then associated with each training instance the predictions made on it by each ensemble member, thus creating a new feature space where the variables are the predictions. They then performed hierarchical clustering on the instances, thus building three subpopulations: one where the prediction accuracy is high, one where it is intermediate and one where it is low [22]. With this approach, Zhanga and Kodell split the original dataset into subpopulations that are easy or difficult to classify [22]. The method seems appealing in general, but does not look promising for the SHIP data: we investigate a three-class problem with a very skewed distribution, so we already know that low accuracy is partially caused by the skew. Hence, we study the dataset exploratively *before* classification, to identify subpopulations that exhibit less skew, and exploratively *after* classification, to identify variables inside each subpopulation, which are associated to the outcome with high likelihood.

Classification Rule Mining. Pinheiro et al. performed association rule discovery on patients with liver carcinoma [23]. The authors pointed out that early detection of liver cancer may help reducing the five-year mortality rate (which is currently 86% [23]), but early detection is difficult, because in the onset of a liver carcinoma, the patient often observes no symptoms [23]. Pinheiro et al. leveraged the association rule miner FP-growth [24] to discover high-confidence association rules and high-confidence classification rules with respect to mortality in a liver cancer patients dataset. We also consider association rules promising for the analysis of medical data, because they are easy to compute and deliver results that are understandable by humans. Therefore, we also use association rules

as baseline mining method, though for epidemiological data and for classification rather than mortality prediction. To use association rules for classification, we specify that the rule consequent should be the target variable.

Clustering. Clustering algorithms are applied to a variety of application domains. For example, clustering is traditionally used for the identification of patient groups, who apply pressure on their feet in a similar way [25]–[28], whereby k-means is the typically the algorithm of choice. In [28], Giacomozzi and Martelli studied the plantar pressure curves of patients with diabetic foot and of control persons, and clustered them on the similarity of the curves with respect to shape and amplitude. They juxtaposed the clusters with respect to predefined groups: for instance, all persons of the group with increased peak pressure and muscle weakness and/or limited mobility of the joints were in one cluster [28]. However, each cluster contained persons from many groups. Deschamps et al. recorded walking trials for patients and controls and studied the relative regional impulses recorded in six forefoot regions [27]: similarly to [28], they also found one cluster that consists only of diabetic patients. Other scholars focused on patients only and strived to understand the variability of pressure distributions among patients and to stratify the patients on plantar pressure similarity. De Cock et al. studied the pressure distributions recorded during jogging, whereby they concentrated on recordings of the six forefoot regions [26]: they used k-means with k=4 and then compared differences in peak pressure and in regional impulse among the identified clusters. Bennetts et al. studied how patients applied pressure when they stood, and performed clustering on peak pressure distinguishing among seven plantar regions [25]. They compared the clusters on mean peak pressure and studied how the number of clusters k affected the results.

2.2 The SHIP dataset

2.3 Subpopulation discovery workflow and interactive mining assistant

2.4 Experiments and findings

2.5 Benefits of our workflow

Chapter 3

Identifying diverse subpopulations

Based on: [12]

3.1 Motivation and comparison with related work

3.2 Finding diverse classification rules

3.3 Experiments and findings

3.4 Benefits our our method

Part II

EXPLOITING TEMPORAL INFORMATION

Chapter 4

Constructing evolution features to capture change over time

Based on: [9]

4.1 Motivation and comparison with related work

Exploitation of temporal information in medicine and healthcare is often observed, foremostly in the analysis of patient records in healthcare applications. For example, Pechenizkiy et al. analyzed streams of recordings to predict whether a patient will need a re-hospitalization after a health failure treatment [29]. Patient monitoring was also the application area of [30], where the emphasis was on computing the similarity between streams of patients for prediction. Combi et al. reported both on the study of streams, e.g. on streams of life signals, and on the temporal analysis of the timestamped medical records of hospital patients [31]. Patient evolution with clustering was studied in [32]: Siddiqui et al. proposed a method that predicts the evolution of a patient from timestamped data by clustering them on similarity and predicting cluster movement in the multi-dimensional space. However, the patient data considered in [32] are labeled at each moment.

A general approach of extracting temporal information is to augment the feature space with variables capturing how individuals *evolve* over time [33]–[35]. These approaches include deriving summary statistics (mean, standard deviation, etc.) from the whole time-series, cutting time-series into small windows before computing aggregations, building separate models for each of the windows which are later combined to an ensemble classifier, or generating shapelets to improve classification accuracy.
For example, Karlsson et al. introduced a framework for the extraction of shapelets from (medical) time-series which are fed into a random forest classifier that uses shapelets as attribute-test conditions instead of regular features [36]. While these approaches are promising for medical time-series with a large number of observations for each individual and only numerical features, they are not applicable for our SHIP and Dan-MONICA data which comprise only three measurement time-points, contain both continuous and categorical features and where the outcome is measured only in the last wave.

There is few work on using temporal information in longitudinal cohort study data with less than 10 waves, although insights from these models could contribute to identify subpopulations that exhibit early signs for an emerging disease, so that timely countermeasures could be initiated. However, with the increasing popularity of deep learning and its many hailed success stories, an alternative to the tedious feature engineering could be the application of deep neuronal networks which automatically construct multiple layers of meta-features that reflect complex non-linear interactions between input features from multiple time points. While these methods often outperform more traditional approaches, they produce so called *black-box models* which are harder to interpret and communicate to a medical expert. For example, in handwriting and speech recognition, it is not important to understand *why* a digit is identified as a *2* rather than a *7* or an audio sample is recognized as *tail* instead of *tale*, as

long as the model is accurate. In contrast, for medical applications it is crucial to understand the reasoning behind a model in order to trust its decision which can have substantial impact on the patient's life quality.

4.2 Overview of the mining workflow

4.3 Generating evolution features

4.4 Experiments and findings

4.5 Benefits our our method

Chapter 5

Feature extraction from short temporal sequences for clustering

5.1 Motivation and comparison with related work

Based on: [2], [3], [11]

5.2 The diabatic foot dataset

5.3 Overview of the mining workflow

5.4 Experiments and findings

5.5 Benefits our our method

Part III

POST-MINING FOR INTERPRETATION

Chapter 6

Post-hoc interpretation of classification models

Based on: [4], [14], [37], [38]

6.1 Motivation and comparison with related work

For medical applications, it is essential for the epidemiological expert to be able to identify the features and feature-value pairs that are most important with respect to the model decision. Gaining actionable insights of a model beyond accuracy is important in order to derive hypotheses on the interaction between potential risk factors and a target outcome.

A few of the widely used algorithms are *intrinsic interpretable* [39], including linear regression, logistic regression and decision trees. In linear regression, the beta coefficient of a predictor variable denotes the difference in the predicted value of the response variable for an increase of 1 unit of the predictor, given all other predictors stay the same. Similarly, the coefficients of a logistic regression model can be interpreted in terms of change of odds for the positive outcome of the response, again given all other predictors stay the same. Decision trees can achieve higher accuracy than linear and logistic regression, in particular for axes-parallel decision boundaries. The tree structure of the model allows for an intuitive visualization and by transposing interpretable classification rules in form of {IF $x \leq v1$ AND $y > v_2$ THEN outcome = positive} can be extracted. However, intrinsically interpretable models are often too simple to deal with complex non-linear feature interactions. Thus, more sophisticated algorithms yielding higher accuracy have become more popular, including support vector machines [40], gradient boosted trees [41] and, in particular, deep neural networks [42]. The superior classification performance comes with a cost though: they lack internal mechanisms for direct model interpretation. To provide a trade-off between high accuracy and model transparency, methods for *post-hoc* explanations and visualizations have been proposed as remedy. One of the most popular *model-specific* interpretation method is the random forest feature importance which makes use of the instances not sampled for the induction of a single tree, the out-of-bag (OOB) instances [43]. First, the model error is measured on the OOB data. Then, the values of each predictor in the OOB data are randomly shuffled at a time and the model error is measured again. The assumption is that the more important a predictor is for the model, the more the error will increase if its values are shuffled. Finally, the difference between the two errors is calculated and averaged over all trees in the forest to yield a feature importance ranking. For neural networks, several scores exist that aggregate the connection weights of all internal and output units that originate from a specific input characterizing a predictor variable [44].

Model-agnostic methods have been proposed to provide algorithm-independent post-hoc explanations and hence, a trade-off between high accuracy and model interpretability. LIME [45] is a framework for local model-agnostic explanations of an arbitrary classifier. For example, if a medical doctor uses a model to classify a patient of interest as low-risk or high-risk patient, he might be not only interested in the classifier decision but also in identifying which

characteristics contributed most to the decision, i.e., whether laboratory values, socio-demographics or rather genetic markers have highest impact. For a complex black-box model, LIME learns a surrogate model to find explanations for the decision of the model for a specific instance of interest i. A surrogate model is an interpretable model, e.g. a linear model or a decision tree, that is used to derive a ranking of feature value importance w.r.t. the classification decision of the model on i. The procedure of LIME is as follows: First, a new dataset is created by random sampling from the training set weighted by the proximity to the instance of interest. The new dataset is applied on the complex model to obtain predictions. Then, the decision boundary of the local dataset is approximated by a simple model. The authors suggest Lasso regression to find the k most important features for the local model.

6.2 The aneurysm dataset

6.3 The tinnitus dataset

6.4 Overview of the mining workflow

6.5 Experiments and findings on aneurysm data

6.6 Experiments and findings on tinnitus data

6.7 Benefits our our method

Part IV

SUMMARY

Chapter 7

Conclusion and future work

7.1 Research results for medical expert-guided knowledge discovery

7.2 Future work

Part V

APPENDIX

Bibliography

[1] H. Völzke, D. Alte, C. O. Schmidt, D. Radke, *et al.*, "Cohort profile: The Study of Health in Pomerania," *International Journal of Epidemiology*, vol. 40, pp. 294–307, 2011. DOI: 10.1093/ije/dyp394.

[2] U. Niemann, M. Spiliopoulou, T. Szczepanski, F. Samland, J. Grützner, D. Senk, A. Ming, J. Kellersmann, J. Malanowski, S. Klose, and P. R. Mertens, "Comparative Clustering of Plantar Pressure Distributions in Diabetics with Polyneuropathy May Be Applied to Reveal Inappropriate Biomechanical Stress," *PLOS ONE*, vol. 11, no. 8, pp. 1–12, 2016. DOI: 10.1371/journal.pone.0161326. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0161326.

[3] U. Niemann, M. Spiliopoulou, J. Malanowski, J. Kellersmann, T. Szczepanski, S. Klose, E. Dedonaki, I. Walter, A. Ming, and P. R. Mertens, "Plantar temperatures in stance position: A comparative study with healthy volunteers and diabetes patients diagnosed with sensoric neuropathy," *EBioMedicine*, vol. 54, p. 102 712, 2020, ISSN: 2352-3964. DOI: https://doi.org/10.1016/j.ebiom.2020.102712. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352396420300876.

[4] U. Niemann, P. Berg, A. Niemann, O. Beuing, B. Preim, M. Spiliopoulou, and S. Saalfeld, "Rupture Status Classification of Intracranial Aneurysms Using Morphological Parameters," in *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, 2018, pp. 48–53. DOI: 10.1109/CBMS.2018.00016.

[5] J. Tuomilehto and K. Kuulasmaa, "WHO MONICA Project: assessing CHD mortality and morbidity," *International Journal of Epidemiology*, vol. 18, no. 3 Suppl 1, pp. 38–45, 1989.

[6] H. Brønnum-Hansen, T. Jørgensen, M. Davidsen, M. Madsen, M. Osler, L. U. Gerdes, and M. Schroll, "Survival and cause of death after myocardial infarction: The Danish MONICA study," *Journal of Clinical Epidemiology*, vol. 54, no. 12, pp. 1244–1250, 2001.

[7] U. Niemann, H. Völzke, J.-P. Kühn, and M. Spiliopoulou, "Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5405–5415, 2014, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2014.02.040. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741400116X.

[8] U. Niemann, M. Spiliopoulou, H. Völzke, and J.-P. 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.

[9] U. Niemann, T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn, "Can we classify the participants of a longitudinal epidemiological study from their previous evolution?" In *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, 2015, pp. 121–126. DOI: 10.1109/CBMS.2015.12.

[10] U. Niemann, T. Hielscher, H. Völzke, T. Ittermann, T. Skaaby, and M. Spiliopoulou, "Exploiting the evolution of study participants to improve classification in longitudinal epidemiological data," *Scientific Reports. To be submitted.*, 2020.

[11] U. Niemann, M. Spiliopoulou, F. Samland, T. Szczepanski, J. Grützner, A. Ming, J. Kellersmann, J. Malanowski, S. Klose, and P. R. Mertens, "Learning Pressure Patterns for Patients with Diabetic Foot Syndrome," in *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, 2016, pp. 54–59. DOI: 10.1109/CBMS.2016.31.

[12] U. Niemann, M. Spiliopoulou, B. Preim, T. Ittermann, and H. Völzke, "Combining Subgroup Discovery and Clustering to Identify Diverse Subpopulations in Cohort Study Data," in *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 582–587. DOI: 10.1109/CBMS.2017.15.

[13] U. Niemann, M. Spiliopoulou, H. Völzke, and J.-P. Kühn, "Subpopulation Discovery in Epidemiological Data with Subspace Clustering," *Foundations of Computing and Decision Sciences (FCDS)*, vol. 39, no. 4, pp. 271–300, 2014. DOI: 10.2478/fcds-2014-0015.

[14] U. Niemann, P. Brueggemann, B. Boecking, B. Mazurek, and M. Spiliopoulou, "Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and socio-demographics," *Scientific Reports*, vol. 10, no. 1, p. 4664, 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-61593-z. [Online]. Available: https://doi.org/10.1038/s41598-020-61593-z.

[15] T. Ittermann, R. Haring, H. Wallaschofski, (...), H. E. Meyer zu Schwabedissen, D. Rosskopf, and H. Völzke, "Inverse Association Between Serum Free Thyroxine Levels and Hepatic Steatosis: Results From the Study of Health in Pomerania," *Thyroid*, vol. 22, no. 6, pp. 568–574, 2012. DOI: http://dx.doi.org/10.1089/thy.2011.0279.

[16] K. Lau, R. Lorbeer, R. Haring, (...), U. John, S. E. Baumeister, and H. Völzke, "The association between fatty liver disease and blood pressure in a population-based prospective cohort study," *Journal of Hypertension*, vol. 28, no. 9, pp. 1829–1835, 2010. DOI: http://dx.doi.org/10.1097/HJH.0b013e32833c211b.

[17] F. Stickel, S. Buch, K. Lau, (...), N. Wodarz, H. Völzke, and J. Hampe, "Genetic variation in the PNPLA3 gene is associated with alcoholic liver injury in caucasians," *Hepatology*, vol. 53, no. 1, pp. 86–95, 2011. DOI: http://dx.doi.org/10.1002/hep.24017.

[18] G. Targher, C. P. Day, and E. Bonora, "Risk of Cardiovascular Disease in Patients with Nonalcoholic Fatty Liver Disease," *New England Journal of Medicine*, vol. 363, no. 14, pp. 1341–1350, 2010. DOI: http://dx.doi.org/10.1056/NEJMra0912063.

[19] M. R. P. Markus, S. E. Baumeister, J. Stritzke, (...), H. Wallaschofski, H. Völzke, and W. Lieb, "Hepatic Steatosis Is Associated With Aortic Valve Sclerosis in the General Population: The Study of Health in Pomerania (SHIP)," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 33, no. 7, pp. 1690–1695, 2013. DOI: http://dx.doi.org/10.1161/ATVBAHA.112.300556.

[20] A. D. Hingorani, D. A. van der Windt, R. D. Riley, (...), W. Sauerbrei, D. G. Altman, and H. Hemingway, "Prognosis research strategy (PROGRESS) 4: Stratified medicine research," *BMJ: British Medical Journal*, vol. 346, no. e5793, 9 pages, 2013. DOI: http://dx.doi.org/10.1136/bmj.e5793.

[21] H. Völzke, G. Fung, T. Ittermann, (...), R. Rettig, B. Rao, and H. K. Kroemer, "A new, accurate predictive model for incident hypertension," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 31, no. 11, pp. 2142–2150, 2013. DOI: http://dx.doi.org/10.1097/HJH.0b013e328364a16d.

[22] C. Zhanga and R. L. Kodell, "Subpopulation-specific confidence designation for more informative biomedical classification," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 155–163, 2013. DOI: http://dx.doi.org/10.1016/j.artmed.2013.04.008.

[23] F. Pinheiro, M.-H. Kuo, A. Thomo, and J. Barnett, "Extracting association rules from liver cancer data using the FP-growth algorithm," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*, 2013. DOI: http://dx.doi.org/10.1109/ICCABS.2013.6629208.

[24] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1–12, 2000.

[25] C. J. Bennetts, T. M. Owings, A. Erdemir, G. Botek, and P. R. Cavanagh, "Clustering and classification of regional peak plantar pressures of diabetic feet," *Journal of biomechanics*, vol. 46, no. 1, pp. 19–25, 2013.

[26] A. De Cock, T. Willems, E. Witvrouw, J. Vanrenterghem, and D. De Clercq, "A functional foot type classification with cluster analysis based on plantar pressure distribution during jogging," *Gait & posture*, vol. 23, no. 3, pp. 339–347, 2006.

[27] K. Deschamps, G. A. Matricali, P. Roosen, K. Desloovere, H. Bruyninckx, P. Spaepen, F. Nobels, J. Tits, M. Flour, and F. Staes, "Classification of forefoot plantar pressure distribution in persons with diabetes: A novel perspective for the mechanical management of diabetic foot?" *PLOS ONE*, vol. 8, no. 11, e79924, 2013.

[28] C. Giacomozzi and F. Martelli, "Peak pressure curve: An effective parameter for early detection of foot functional impairments in diabetic patients," *Gait & posture*, vol. 23, no. 4, pp. 464–470, 2006.

[29] M. Pechenizkiy, E. Vasilyeva, I. Zliobaite, A. Tesanovic, and G. Manev, "Heart Failure Hospitalization Prediction in Remote Patient Management Systems," in *Proc. of IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, 2010, pp. 44–49. DOI: http://dx.doi.org/10.1109/CBMS.2010.6042612.

[30] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, "A System for Mining Temporal Physiological Data Streams for Advanced Prognostic Decision Support," in *IEEE Int. Conf. on Data Mining (ICDM)*, 2010, pp. 1061–1066. DOI: 10.1109/ICDM.2010.102.

[31] C. Combi, E. Keravnou-Papailiou, and Y. Shahar, *Temporal Information Systems in Medicine*. Springer, 2010.

[32] Z. F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu, "Predicting the post-treatment recovery of the patients suffering from TBI," in *Int. Conf. on Brain Informatics and Health (BIH'14)*, Warsaw, Poland, Aug. 2014.

[33] K. Orphanou, A. Stassopoulou, and E. Keravnou, "Temporal abstraction and temporal Bayesian networks in clinical domains: A survey," *Artificial Intelligence in Medicine*, vol. 60, no. 3, pp. 133–149, 2014.

[34] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," *Journal of Biomedical Informatics*, vol. 53, pp. 220–228, 2015.

[35] J. Zhao, P. Papapetrou, L. Asker, and H. Boström, "Learning from heterogeneous temporal data in electronic health records," *Journal of Biomedical Informatics*, vol. 65, pp. 105–119, 2017.

[36] I. Karlsson, P. Papapetrou, and H. Boström, "Generalized random shapelet forests," *Data Mining and Knowledge Discovery*, vol. 30, no. 5, pp. 1053–1085, 2016.

[37] U. Niemann, B. Boecking, P. Brueggemann, W. Mebus, B. Mazurek, and M. Spiliopoulou, "Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables," *PLOS ONE*, vol. 15, no. 1, pp. 1–18, Jan. 2020. DOI: 10.1371/journal.pone.0228037. [Online]. Available: https://doi.org/10.1371/journal.pone.0228037.

[38] U. Niemann, B. Boecking, P. Brueggemann, B. Mazurek, and M. Spiliopoulou, "Gender-Specific Differences in Patients With Chronic Tinnitus—Baseline Characteristics and Treatment Effects," *Frontiers in Neuroscience*, vol. 14, p. 487, 2020, ISSN: 1662-453X. DOI: 10.3389/fnins.2020.00487. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2020.00487.

[39] Z. C. Lipton, "The Mythos of Model Interpretability," in *IEEE ICML Workshop on Human Interpretability of Machine Learning*, 2015, pp. 121–126.

[40] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. of Workshop on Computational Learning Theory*, ACM, 1992, pp. 144–152.

[41] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[43] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[44] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological Modelling*, vol. 178, no. 3-4, pp. 389–397, 2004.

[45] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proc. of ACM SIGKDD Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.