Doctoral theses at NTNU, 2022:390

Deepika Verma

Using Case-Based Reasoning for Creating Intelligent Systems in Healthcare

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Deepika Verma

Using Case-Based Reasoning for Creating Intelligent Systems in Healthcare

Thesis for the Degree of Philosophiae Doctor

Trondheim, December 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Computer Science

© Deepika Verma

ISBN 978-82-326-6543-3 (printed ver.)
ISBN 978-82-326-5426-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:390

Printed by NTNU Grafisk senter

Alt til sin tid ...

Abstract

Healthcare research is an emerging field of application of machine learning techniques to investigate complex health datasets. *Patients* are at the center of any healthcare system. There is a growing realisation of the patient-centred healthcare system and the notion is slowly changing the healthcare scenario from a *one glove fits all approach* to a more *personalised approach*. Data collected in population-based and intervention-based studies has immense potential in supporting primary caregivers in providing patient-centred care by facilitating clinical decision-making. From identifying patients at a high risk of post-surgical complications to forecasting their quality of life, recent developments in healthcare informatics suggest that leveraging the capabilities of machine learning techniques on complex health data can have a significant impact on the decision-making process in clinical settings. To this end, we explored supervised and unsupervised machine learning methods, predominantly, Case-Based Reasoning (CBR) methodology.

The overall theme of this research is exploring the potential of healthcare datasets using CBR methodology. We used two unique and innovative datasets—a population-based dataset consisting of objectively measured physical behaviour data collected using body-worn sensors in HUNT4 cohort study, and an intervention-based dataset comprising patient-reported outcome measurements collected during clinical trials to test the efficacy of tailored interventions in SELFBACK mobile app—and applied both supervised and unsupervised learning to glean valuable information.The goal is to develop intelligent modules that can be incorporated into clinical decision support systems to support clinicians in the informed decision-making process or as standalone systems. Focus is placed on applying the case-based methodology to learn from the data without making assumptions. The HUNT4 physical behaviour dataset was investigated to get insights

into the physical behaviour characteristics of the population and identify clusters of similar behaviour profiles using a new clustering approach. The clustering approach was valuable in identifying groups of similar physical behaviour, which can be used further by primary caregivers to underpin the amount of physical activity tailored to the individual's needs. The SELF-BACK intervention datasets were explored to determine the predictors of various patient-reported outcomes and investigate the predictive potential of the patient-reported outcome measurements using case-based and conventional machine learning methods. The methods used show the potential to predict pain-related patient-reported outcomes.

Overall, our results indicate that a close liaison between healthcare data, clinicians, and machine learning methods can promote a better understanding of achieving patient-centred care through the addition of intelligent systems in clinical decision support. The results also provide grounds for further research and development of evidence-based clinical decision support systems.

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* in Computer Science at the Department of Computer Science, Norwegian University of Science and Technology (NTNU). The research presented here was conducted under the supervision of Professor Kerstin Bach and co-supervision of Paul Jarle Mork.

This thesis takes the form of a paper collection where the included papers have either been published or submitted at conferences or in journals. The papers have been reformatted and typeset anew for inclusion in this thesis for consistency and readability. For this reason, the papers deviate visually from their published or submitted versions.

Acknowledgements

First, I would like thank my supervisors Professor Kerstin Bach and Professor Paul Jarle Mork for providing me this opportunity to pursue this PhD degree and guiding me in the process. I am grateful for your invaluable support, supervision and inspiration through the entire journey.

A special thanks to Professor Anders Kofod-Petersen for his guidance during the doctoral consortium at the *International Conference on Case-Based Reasoning 2019*.

I wish to thank my colleagues Amar Jaiswal, Shweta Tiwari, Sylvester Sabathiel, Ilya Ashikhmin, Biswanath Barik, Rabail Tahir, Daniel Groos, Bjørn Magnus Mathisen, Sverre Herland and many more for sharing their knowledge and for interesting discussions. A special thanks to Randi Holvik and Kai Dragland for your kind words and advice. I have always enjoyed our conversations in the office and found your words soothing.

I would also like to thank my friends who have supported me through these years. I am grateful for your words of encouragement and kindness.

Finally, and most importantly, I want to express my deepest, heartfelt, and most sincere gratitude to my mother Narinder Verma, father Shashi Kant Verma, and sister Kim Verma for always standing by me and supporting me throughout this journey. You have always encouraged me to pursue my career goals and given me the strength and confidence to do so. Thank you for believing in me. I would not have been able to finish this journey without your love and support. I would also like to thank my partner Andreas T.G. Janssønn for his constant support, love, and encouragement. Thank you for bearing with me through the most challenging phase of my life as much as our lives, for listening patiently, and for being kind to me.

Contents

List of Figures

List of Tables

Part I

RESEARCH OVERVIEW

Chapter 1

Introduction

1.1 Context and Motivation

Increased use of information systems in health services and collection of patient-specific information generates large amounts of healthcare data. The data collected has immense potential for the provision of healthcare services as well as quality improvement, research, public health, management, and planning. The combination of colossal amounts of varied data, increased computing power owing to the technological advancements, and the use of intelligent methods provides a golden opportunity for analysing complex health data to get deeper insights and build prognostic models that can be incorporated into clinical decision support systems to facilitate informed decision-making. Machine learning algorithms are increasingly being used in various areas of healthcare research (Quazi, 2022; Krishnamoorthi et al., 2022; Sanchez et al., 2022; Fryan et al., 2022). Technological developments have begun having an impact on the public healthcare ecosystem but are yet to show their full potential (Louw et al., 2017; Sarwar et al., 2018; Alanazi, 2022). These developments serve to improve the overall clinical workflow from the preventive and diagnostic phase to the prescriptive and restorative phase in healthcare. Healthcare data has immense potential in accelerating the growth of clinical research, both objective and subjective measurements provide an avenue for furthering patient-centred care with the assistance of tools that can facilitate utilising their potential and supporting clinical decision-making (Jensen et al., 2012; Wang and Gottumukkala, 2021). As the healthcare sector becomes more proactive in using machine learning

techniques, more research and development are necessitated to realise the full potential of the technology for complex healthcare datasets.

Physical inactivity, a growing public health concern worldwide, is estimated to be responsible for about 9% of premature mortality (Lee et al., 2012) and contributes to a wide variety of chronic diseases and conditions including low back pain, cardiovascular diseases, stroke, type 2 diabetes, and cancer (Picavet and Schuit, 2003; Fox, 2004; Lynch and Leitzmann, 2017). Health surveys globally reveal large fractions of the population not meeting the physical activity levels recommended by the world health organisation (WHO) (Guthold et al., 2008). To this end, activity recommendation systems have the potential to promote physical activity and thereby contribute to the improvement of public health (Smyth, 2019). With the growth in popularity of wearable activity trackers, objective physical behaviour measurements in population-based studies have opened new possibilities for public health and computer science researchers alike. Body-worn sensors have helped in making the shift from self-reported physical activity data to objectively measured physical activity data, thereby eliminating the bias due to self-reporting. The use of machine learning methods has further simplified the extraction of and utilization of raw sensor data (Arif and Kattan, 2015). Population-wide cohort studies collecting objective physical behaviour measurements, such as the HUNT4 [1] in Norway, provide an opportunity for assessing detailed physical behaviour patterns. Investigating determinants of physical activity behaviour can inform the development of interventions aimed at improving physical activity level. An important step here would be to identify groups with similar physical behaviour profiles (Marschollek, 2013; Kohl 3rd et al., 2012).

Non-specific neck pain (NP) and low back pain (LBP) are another public health concern and a leading cause of disability worldwide (Hurwitz et al., 2018). Almost all the reported cases of neck and or low back pain (NLBP) are non-specific, meaning that they cannot be attributed to any specific cause such as a disease, infection, malignancy or fracture (Hartvigsen et al., 2018), but have a high recurrence frequency (Côté et al., 2004; Andersson, 1999). As many as 70-80% of all adults experience LBP and 20-70% experience neck pain (NP) at some point in their life (Bovim et al., 1994; Brattberg et al., 1989; Kelsey et al., 1980; Sinnott et al., 2017). These pain conditions are the main cause of early retirement and are responsible for the great-

[1] *www.ntnu.no/hunt/hunt*4

est loss of productive life years in the workforce compared with other non-communicable diseases (Briggs et al., 2018). Considering the vast impact on individual well-being and public health, there is a need to develop scalable and cost-effective interventions to improve outcomes in people with LBP and NP. Designing such interventions requires insight into modifiable factors, that are known to influence the prognosis of symptoms and allow for the prediction of outcomes. Patient-reported outcome measurements (PROMs) collected routinely in clinical settings to evaluate pain-related symptoms of patients with LBP or NP provide an opportunity to examine short to long-term predictors of outcomes that can support informed decision-making (Baumhauer, 2017). PROMs are being increasingly given more importance than any other outcomes like clinical, physiological, or clinician-reported (Wang and Gottumukkala, 2021). Research indicates enhanced treatment adherence and outcomes can be obtained by giving attention to patient feedback on healthcare outcomes and patient behaviour change (Carroll, 2002).

Literature suggests leveraging machine learning techniques for healthcare datasets can have a meaningful impact on the further development of clinical decision support systems (CDSS) (Panch et al., 2018). Machine learning methods are increasingly being applied in clinical studies to investigate complex healthcare datasets and predict outcomes such as comorbidities, drug efficacy, patient stratification, and quality of life among others (Shi et al., 2012; Moonesinghe et al., 2013). *Patients* are at the center of any healthcare system and there is a growing realisation of the need for a patient-centred healthcare system (Louw et al., 2017). Predictive analytics is expected to play a key role for prevention of diseases at both individual and population-wide levels. Enabling the use of prognostic analytics on patient-centred data can support caregivers and other involved parties to dispense targeted interventions to prevent the occurrence of a worse clinical outcome or improve physical activity levels.

Machine learning methods provide a promising approach to explore complex health datasets (Adkins, 2017) and build models that can learn from the data and generalise to facilitate evidence-based decision-making in clinical practice (Vasquez-Morales et al., 2019). Case-based reasoning (CBR) is a machine learning approach with a rapidly growing field of research and development within healthcare informatics and broad applicability to building intelligent systems in health sciences domains (Bichindaritz et al., 2008; Bichindaritz and Marling, 2010). CBR has been demonstrated to be a suitable methodology to apply in unstructured domains such as multidisci-

plinary medical services (Chuang, 2011). The value of CBR stems from capturing specific clinical experience and leveraging this contextual, instance-based, knowledge for solving clinical problems. CBR systems offer means of abstracting and transferring specific domain expert knowledge into a self-explanatory and user-friendly tool, which can be used to generate explainable solutions for problems ranging from simple daily life tasks to complex issues (Weber et al., 2005; Vasquez-Morales et al., 2019).

CBR has been widely applied for classification of medical data (Yao and Li, 2010; Campillo-Gimenez et al., 2013), physical activity data (Uddin and Loutfi, 2013) and has also proven useful in clinical practice for decision support, explanation, and quality control (Holt et al., 2005). To continue this line of research in this doctoral work, we focused on data-driven research to address different aspects of utilisation of healthcare datasets to facilitate advancing the application of machine learning methods on these datasets. Using CBR, we implement learning models that can be incorporated in decision support systems or public health research to discover new knowledge or can even be used standalone to find new information. We utilised two datasets in our research work—the first consisting of objective measurements of physical behaviour from a population-based cohort study called HUNT4 and the second comprising subjective clinical measurements reported by patients with non-specific LBP or NP from intervention-based clinical trials for the SELFBACK project. With new ways of capturing data, these datasets present new challenges concerning their utilisation for creating intelligent systems that can add value to the current healthcare system. HUNT4 provides a more unbiased setting for understanding the physical activity aspect of a small population from an objective point of view. While the subjective setting in SELFBACK , where we have an innovative intervention to address a chronic health issue (LBP or NP), provides an avenue to understand whether the intervention is helpful for the patients from the patient's perspective. Such unique healthcare datasets have immense potential to improve the quality of healthcare research and healthcare services delivered to individuals. To this end, we used CBR to address several aspects of developing intelligent modules to better utilise healthcare datasets including the development of similarity measures for CBR modelling in a data-driven manner, utilisation of the data-driven similarities to cluster case bases in CBR, and investigating approaches for selecting important clinical measurements from subjective healthcare datasets that can facilitate prediction of patient-specific outcomes. We used conventional machine learning

methods in addition to compare the performance of the CBR models.

1.2 Research Questions

This section details the overarching objectives of this research. The research goals are stated first, followed by three specific research questions. The starting point for this doctoral research was to better utilize healthcare datasets for facilitating informed clinical decision-making. Are these datasets sufficiently understood and utilized that we can begin using them for decision making? Can we learn new, valuable insights by leveraging the learning abilities of machine learning methods? From a clinical point of view, a holistic perspective of how the different patient-centred data collected routinely can be incorporated into a decision-making tool that allows one to identify population-wide similar behavior, follow the trajectory of the patients, predict individual outcomes, and support informed decision-making. From a computer science standpoint, the clinical perspective allows us to investigate how the knowledge in a healthcare dataset may be represented in machine learning models and how the models may be utilized to support informed decision-making. The research goals and research questions in the thesis are motivated by these overarching perspectives.

> **Research Goal**:
> Advancing research within application of machine learning methods on healthcare datasets and developing methods that can be used to further the development of CDSS.

Addressing the research goal will further the understanding of what the domain needs and push the boundaries to incorporate intelligent methods into data-driven research and build evidence-based clinical decision-making tools. To address the research goal, the work in this doctoral research has been split into three research questions.

> **Research Question 1**:
> How to measure similarity among different individuals based on their objective and subjective measurements?

The first research question addresses the development of a methodology to build suitable similarity measures that represent the similarity between

two given patients in any given healthcare dataset. The methodology should be applicable on both objective and subjective measurements in any healthcare (or another domain's) dataset.

> **Research Question 2**:
> How can machine learning methods be applied to subjective patient-reported datasets to facilitate individualized patient-reported outcome predictions?

The second research question addresses the utility of machine learning methods in analysing subjective healthcare datasets and investigates whether the case-based methodology can be useful in building prediction models that can make individualized predictions from the given healthcare data.

> **Research Question 3**:
> What are the state-of-the-art of machine learning methods for investigating patient-reported outcome measurement datasets?

When pursuing the second research question, it becomes natural to look at the state-of-the-art. Therefore, the third and final research question looks into the existing literature that involves the application of machine learning methods to investigate healthcare datasets comprising (subjective) patient-reported measurements.

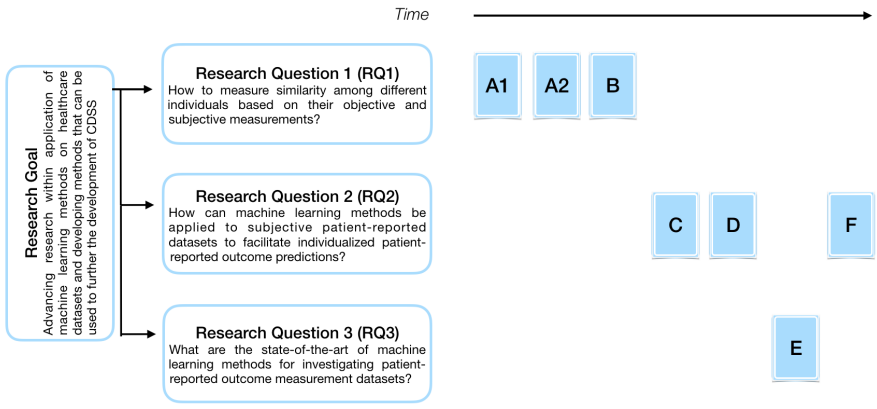


Figure 1.1: Overview of the research conducted as a part of this thesis.

1.3 Overview of the Research Conducted

This section provides an overview of the publications included in this thesis. Seven papers, labeled A-F are included in this thesis and they can be found in their entirety in the second part of this thesis. Figure 1.1 gives a visual representation of the research questions and how the research papers relate to them.

The first paper, A1, addresses the RQ1 by introducing a data-driven approach for local similarity modelling and demonstrating the similarity estimation amongst individuals based on their objective physical activity measurements. This paper is published in the proceedings of the *International Conference on Case-Based Reasoning (ICCBR 2018)* .

- **Paper A1**: Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Modelling Similarity for Comparing Physical Activity Profiles - A Data-driven Approach. In Michael T. Cox, Peter Funk, and Shahina Begum, editors, *International Conference on Case-Based Reasoning*, pages 415-430, Cham, 2018. Springer. ISBN 978-3-030-01081-2

The methodology introduced in this conference publication is later published as a position paper in the 2019 symposium of the *Norwegian AI Society* (*NAIS 2019*). The position paper, Paper A2, includes the methodology and background from the original paper A1 and extends on it by demonstrating the validity of the proposed method on other datasets. The paper underwent peer-review before being accepted and published.

- **Paper A2**: Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Similarity Measure Development for Case-Based Reasoning– A Data-driven Approach. In Kerstin Bach and Massimiliano Ruocco, editors, *Norwegian Artificial Intelligence Society*, pages 143–148, Cham, 2019. Springer. ISBN 978-3-030-35664-4

The subsequent paper extends on the similarity measures developed in paper A1 and presents a clustering algorithm to cluster case bases using the similarity measure as the clustering metric and thereby supports RQ1. This paper is published in the proceedings of the *International Conference on Agents and Artificial Intelligence (ICAART 2020)*.

- **Paper B**: Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Clustering of Physical Behaviour Profiles Using Knowledge-intensive Similarity Measures. In Ana Rocha, Luc Steels, and Jaap van den Herik, editors, *International Conference on Agents and Artificial Intelligence*, Volume 2, pages 660–667. INSTICC, SciTePress, 2020. ISBN 978-989-758-395-7.

With the next paper, we move our investigation under RQ2 into healthcare datasets with a dataset that consists of subjective clinical measurements or patient-reported measurements. In paper C, we explore the application of a number of machine learning algorithms on subjective patient-centred measurements and develop outcome prediction models. This paper is published in the journal *BMC Medical Informatics and Decision Making*.

- **Paper C**: Exploratory Application of Machine Learning Methods on Patient Reported Data in the Development of Supervised Models for Predicting Outcomes. *BMC Medical Informatics and Decision Making*, 22(227), 2022. ISSN 1472-6947.

Continuing our research into the subjective healthcare datasets under RQ2, we address some shortcomings of the machine learning models presented in Paper C, by looking into other methods of feature selection and build case-based prediction models for predicting individualised patient-reported outcomes. This paper is published at the conference *British Computer Society, Specialist Group on Artificial Intelligence (BCS SGAI 2021)*.

- **Paper D**: Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Using Automated Feature Selection for Building Case-based Reasoning Systems: An Example from Patient-Reported Outcome Measurements. In Max Bramer and Richard Ellis, editors, *British Computer Society, Specialist Group on Artificial Intelligence*, pages 282–295, Cham, 2021. Springer. ISBN 978-3-030-91100-3.

The next paper addresses the RQ3 and presents a literature review to summarise the existing state-of-the-art application of machine learning methods on PROM datasets for predicting individualised patient-reported outcomes. This paper is published in the journal *MDPI Informatics* in late 2021.

- **Paper E**: Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Application of Machine Learning Methods on Patient-Reported Outcome Measurements for Predicting Outcomes: A Literature Review. *MDPI Informatics*, 8(3), 2021. ISSN 2227-9709.

The final paper continues the investigation under RQ2 to address the challenge of external validation of prediction models developed for patient-reported outcomes. This paper extends on the prediction models presented in Paper D and presents an external validation of the models using an external dataset. This paper has been accepted for publication in the *Elsevier International Journal of Medical Informatics*.

- **Paper F**: External Validation of Prediction Models for Patient-Reported Outcome Measurements collected using the SELFBACK Mobile App. *Accepted for publication in Elsevier International Journal of Medical Informatics*

1.4 Thesis Structure

This thesis is composed of two parts and is structured as follows:

- **Part I: Research Overview**
  This part includes the introduction to the research work in chapter 1, background and related work in chapter 2, a comprehensive description of the datasets in chapter 3, an overview of the research methodology in chapter 4, results and evaluation of research questions in chapter 5 and finally, discussion and conclusion of the thesis in chapter 6.

- **Part II: Publications**
  This part contains full-length research papers included in this thesis.

Chapter 2

Background and Related Work

As mentioned in the last chapter, we looked into different aspects of utilising healthcare datasets to facilitate advancement of research in the field of clinical decision support systems and focus on leveraging the analytical capabilities of machine learning methods on healthcare datasets. The first aspect that we look into is development of similarity measures to measure the similarity among patients in a given healthcare dataset, based on either objective measurements such as sensor-based physical behaviour measurements or subjective measurements such as clinical questionnaires. The second aspect that we look into is unsupervised learning for CBR systems, more specifically, clustering a case base while preserving the semantic relationship between the cases. And finally, we investigated various machine learning approaches to determine predictors of clinical outcomes and to forecast patient-specific outcomes. This chapter addresses the background and related work of each aspect and gives a brief introduction to *myCBR*, a similarity-based system development and retrieval tool.

Before continuing further, we would briefly define some pertinent terms in the CBR methodology to facilitate clarity in the sections ahead.

- **Case-Based Reasoning**: CBR is a methodology for intelligent reuse of existing knowledge of already solved problems (called **cases**) to solve new problems (Aamodt and Plaza, 1994). Each case contains a problem description and a corresponding solution that can be used to solve a future problem.

- **Case Base**: A case base is a memory that contains a collection of solved problems as cases.

- **Similarity**: The problem description of a case is said to be *similar* to the new problem if its solution can be useful for adapting a solution for the new problem. The term *similarity* refers to the *utility* of the *similar* case for the new problem by a certain measure. This measure is known as the **similarity measure**.

- **Retrieval**: Retrieval is the process of searching through the case base to find cases that are *similar* to the new problem.

- **Reuse**: CBR aims to generate solutions for new problems by reusing the solutions of similar cases from the case base. If a new problem is the same as a case in the case base, then the solution of this case can directly be reused to solve the new problem. However, if the case is *similar* but not the same, then the solution must be adapted to suit the new problem. This process of adapting the solution is called **adaptation**.

- **Revise and Retain**: Revise aims to evaluate the validity of an adapted solution for a new problem. If the adaptation is valid and revising generates a new case, it may be included or **retained** in the case base.

- **CBR cycle**: The CBR cycle, introduced by Aamodt and Plaza (1994), consists of the four Rs: *Retrieve*, *Reuse*, *Revise* and *Retain*. In the *Retrieve step*, the system searches for a subset of cases from the case base that are similar to the new problem. In the *Reuse* step, the system adapts the solution of the retrieved cases to the new problem. In the *Revise* step, the system evaluates the correctness of the adapted solution. Finally, in the *Retain* step, the system decides whether or not to include the new case in the case base.

2.1 Similarity Measures Development

The similarity measure, used to quantify the degree of resemblance between a pair of cases, plays a central role in the retrieval of similar cases from the case base. This is why CBR systems are also known as *similarity searching systems*. The notion of similarity in CBR is useful for finding past experiences in the knowledge base in order to solve a new problem. The *local-*

global principle for development of similarity measures for a concept X composed of atomic parts z_i, also known as attributes of the concept, emphasises that each concept must be compared locally at the atomic level as well as globally on the concept level (Richter and Weber, 2013). This means that similarity measured should first be constructed at the attribute level for each attribute z_i, known as local similarity measures, and then at the conceptual level for the object X, known as global similarity measure, in order to reflect a global view of the concept. However, similarity is a relative notion rather than an absolute notion, in the sense that similarity is measured relative to some aspect and is not fixed. There can be many possible similarity measures to assess the similarity between the same two experiences, each based on different aspects. Thus, the definition of a similarity measure is also an important part of the entire problem-solving task in CBR. Furthermore, similarity measures can be thought of as a heuristic that is used to estimate the utility of the cases in the case base for solving a particular problem, and therefore must approximate a *utility function*. The implication is that while simple similarity metrics can measure the syntactic similarity between two cases, they may not measure the semantic similarity. This is because semantic similarity captures the domain knowledge while syntactic similarity does not. The use of simple similarities or knowledge-poor similarities may insufficiently approximate the utility of the cases and lead to poor retrieval results (Stahl and Gabel, 2003).

Similarity measures are a crucial component in any CBR system and therefore, a considerable portion of the existing literature on the development of CBR systems focuses on the development of suitable similarity measures. Not just within CBR, learning and adapting similarity measures is also an extensively studied theme in the field of traditional machine learning methods, especially automatic acquisition of similarity measures (Mountrakis et al., 2005; Kang et al., 2017). From the earliest days of research in similarity assessment in CBR systems, the standard methodology had been to assess similarity based on the feature vector representation of the cases using metrics that utilised the feature values. More novel mechanisms of similarity assessment arrived with further research that used strategies other than the established techniques. Cunningham (2009a) gave an overview of some of the novel similarity learning strategies and proposed a taxonomy that organises these new mechanisms in the context of the established techniques. This taxonomy organized the similarity assessment strategies into four categories—direct metrics such

as Euclidean distances, transformation-based such as graph edit distance, information-theoretic such as compression-based distance, and lastly, emergent measures such as web-based or cluster kernels. Stahl et al. have made significant contributions to the field of learning similarity measures (Stahl, 2001, 2005; Stahl and Gabel, 2006). Stahl (2001) proposed a framework he called "*similarity teacher*" that utilises the *teacher* (domain expert) feedback on the quality of retrieval to automatically learn similarity measures. The idea here is that if the *teacher* knows the correct order of cases in the retrieval, this feedback can be used to learn the feature weights and update the similarity measures.

The existing literature often makes a distinction between learning similarities in the problem space versus in the solution space. While most of the existing literature almost exclusively focuses on similarity learning in the problem space, attempts have been made to learn similarities in the solution space. The idea was introduced by Stahl and Gabel (2003) who proposed using an evolution program, a form of genetic algorithms, to learn similarity measures that can sufficiently approximate the utility of cases. The authors later used neural networks to learn the local similarity measures of the attribute and the weights of the modelled similarity measures in global similarity (Stahl and Gabel, 2006). Abdel-Aziz et al. (2014) proposed a learning method that adapts the similarity of the solution based on the gathered experiences in a previously proposed preference-based CBR system (Hüllermeier and Schlegel, 2011), showing that the data distributions and distances in data sets can be used for learning similarity measures. The main idea was to minimize the distance between the ideal solution to a new problem and an existing solution to an existing problem in the case base. While an ideal solution may not exist, the authors use the learning preferences collected in a problem-solving episode to adapt the distance measure using Bayesian learning. What is different in their approach is that they focus on similarity in the solution space, and not in the problem space. The difference lies in the idea that while the generated solution may not be the ideal one, it is expected to be closest to the practically ideal solution. More recently, Mathisen et al. (2020) presented a framework distinguishing four types of similarity measures to facilitate automating the development of similarity measures. The authors analysed the existing similarity measure construction methods and devised two novel designs that utilise machine learning methods to learn similarity measures. One uses a siamese neural network classifier for measuring similarity while the other uses a combination of static binary func-

tions and neural networks to learn similarity measures. Gabel and Godehardt (2015) had earlier proposed a neural network approach to automate the learning of similarity measures by concatenating two data points into a single input vector. The authors made use of a "black-box" approach to learn neural network-based local and global similarity measures in so-called *"similarity clouds"* and used these to induce human-readable and interpretable similarity measures.

2.2 Case-Based Clustering

While supervised learning works on a pre-defined hypothesis about a given dataset, unsupervised learning can be useful to look for patterns and clusters to get insights that add value to the data without any guiding assumptions. In CBR systems, a case base is one of the most important knowledge containers. A case base is, as the name suggests, a collection of cases storing previously solved problems and their solutions. The case base is organized to facilitate retrieval of the most similar cases in the event of the arrival of a new problem. When a new problem arrives, the system searches its case base to look for similar past cases. The solution of the retrieved past case(s) provides a starting point for generating a solution for the new problem. A quality case base is critical for the success of a CBR system since, without the prior problem-solving experiences (cases), the system becomes vain. One of the active research problems in CBR is case organization and retrieval when the case base is large or unlabelled, or when there is a need for diversity in solutions. In such scenarios, partitioning the case base into several clusters is helpful in identifying meaningful patterns, organising the case base in a meaningful way and extracting valuable knowledge from the clusters that can make the case retrieval process more efficient.

Several methods and algorithms for clustering have been introduced for the organisation and maintenance of case bases in CBR systems. Self organizing maps (SOM) have found popularity in many of these methods and algorithms proposed over the years in the developments of CBR systems. SOMs were first introduced by Kohonen (1982) and are a neural network-based tool that can be used for clustering of data and visualization. SOM can uncover hidden semantic relationships in textual data owing to their ability to create spatially organized representations of features in input signals and their abstractions (Kohonen, 1990). Kim and Han (2001) used SOM and an-

other clustering technique, learning vector quantization to produce adaptive clusters in a cluster-indexing method for bond-rating prediction CBR systems. They compute cluster centroids using the two clustering techniques and add the centroids as new artificial cases to represent information in each cluster, and later use these centroids for case recall. Similarly, Zhuang et al. (2009) used SOM to partition a case base consisting of pathology data of patients into several clusters and presented a CBR system that can be used to provide evidence-based decision support to general practitioners regarding ordering pathology tests for new cases. The authors successfully demonstrate that such systems can be useful for clinical decision support by easing the burden on the healthcare provider, stratifying the process based on documented clinical experience and external evidence from systematic research as well as reducing the risk of judgment errors posed by information overload and time constraints on the practitioners. This has also been supported by Van Der Weyden (1999). Despite their outstanding abilities, the traditional SOM suffer drawbacks, namely, they require a pre-defined topology of the network and lack support for visualization of hierarchical clusters. To overcome these shortcomings, Zhu et al. (2015) proposed a growing hierarchical SOM (GHSOM) to partition the initial case base into smaller subsets and organize the subsets into a flexible and hierarchical structure that consists of multiple layers of independent SOM. The GHSOM structure was found to lead to more efficient case retrievals in their case study on ten open source datasets from the UCI machine learning repository. Müller and Bergmann (2014) also presented a hierarchical cluster-based indexing approach for process-oriented CBR systems, where they used the modelled similarity measure to construct a hierarchical cluster tree that acts as an index to improve retrieval of similar cases from subsets of the case base. They found their approach to have higher retrieval quality and lower retrieval time of semantic workflows compared to a linear retrieval due to faster traversal to find similar cases in the cluster tree. Other similar work includes that of Lucca et al. (2018) that presented a framework to organize large case bases into smaller sub-case-bases and developed an index on the clustered sub-case-bases for efficient retrieval of relevant cases in agent simulation systems.

Clustering techniques have often been combined with CBR not only for improving the efficiency of case recall and case base organization and maintenance but also for improving case-based classification, case generation, labelling as well as adaptation (Clerkin et al., 1994; Arshadi and Jurisica,

2005). Wiratunga et al. (2003) presented a framework that utilizes a clustering approach for labelling unlabelled cases . The authors proposed to cluster the unlabelled problems within the case base into smaller subsets using a decision tree index built over the case base, which can then be labelled with appropriate solutions by the domain expert. Cunningham (2009b) had previously suggested using similarity as a measure for selective sampling and generating solutions for unlabelled cases in clustered case bases. Fanoiki et al. (2010) proposed using a similar cluster-based approach to that of Wiratunga et al. (2003) for solution generation by identifying relevant cases for a given query not just in the problem space but also in the solution space. In their work, they formulate a solution by first selecting the cluster with the most similar problem description and then adapting the solution to the cases within that cluster. Arshadi and Jurisica (2005) employed a CBR ensemble approach they call "*mixture of experts*" to predict labels of unseen high-dimension cases in several medical datasets. The authors first employed spectral clustering to partition the dataset into k clusters followed by feature selection for each cluster such that each cluster acts as an independent case base for k CBR experts and the final label is determined based on a gating network that computes a weighted average of the expert votes.

2.3 Machine Learning Methods in Healthcare

The topic of *machine learning in healthcare* is fairly broad and encompasses several fields of application. In this thesis, however, we concern ourselves with machine learning methods that find utility in driving patient-centred research that could support informed clinical decision-making.

Increased use of technology and its integration in how patient data is collected and stored has played an important role in making predictive analytics possible and useful, both for patients as well as healthcare providers (Alharthi, 2018). Personalised treatment and recommendations represent an approach that has the potential to transform patient-centred healthcare (Giga, 2017). CDSS assimilate expert knowledge based on observations from previous patients and use empirical findings to predict outcomes for a new patient, taking into account their past observations (Velickovski et al., 2014). Such a tailored and holistic approach enables capturing the complex interactions of clinical factors pertinent to each patient to provide treatment that suits the patient's individual needs and symptoms (Bitton et al., 2014). Pri-

mary practitioners and clinicians use their experience, clinical guidelines, practices, and professional judgment to determine the most suitable course of action for the patients. Predictive systems in clinical settings could assist in decisions relating to the treatment and its response by assessing the observations of the patient and providing evidence-based information on the best-suited treatment for a given individual since *one glove does not fit everyone* (Sepucha et al., 2018). Both supervised and unsupervised learning techniques provide valid analytical outlets for such applications.

The existing literature has demonstrated the value and efficacy of machine learning methods on EHR datasets, all from scheduling surgeries (Shahabi Kargar et al., 2014; Devi et al., 2012) to predicting the risk of post-surgical mortality (Wong et al., 2017; Moonesinghe et al., 2013; Marufu et al., 2016) among others. Despite the broad application of machine learning methods in the biomedical field, their utilisation in clinical research concerning PROMs for patient-centred care and precision treatment/medicine remains low. The last decade has seen a slow but steady surge in researchers turning towards the inclusion of machine learning methods to delve into the ever-growing patient-reported data and uncover hidden associations that are important in facilitating clinical decision-making (Giga, 2017; Buell, 2016), and exploring the potential of the methods for PROMs and of PROMs for their predictive prowess. Harris et al. (2019) found three machine learning algorithms—gradient boosting machines, logistic regression, and quadratic discriminant analysis—to provide modest results for predicting post-surgical improvement in several patient-reported outcomes based on pre-operative PROMs. Based on their observations, the authors argued for higher integration of such models into shared clinical decision-making to improve patient satisfaction. Fontana et al. (2019) found the prediction models they built to perform modest-to-good (AUC: [0.60,0.89]) depending on the outcome predicted using pre- and post-surgical PROMs collected at four different time-points. The authors also reported that the features collected at two specific time points, i.e., *before decision* and *before surgery* were enough as predictors, and the addition of more information did not lead to an improved model performance. Polce et al. (2020) made similar findings when predicting post-surgical patient satisfaction using a PROMs dataset consisting of sixteen features. Using recursive feature elimination with random forest, the authors arrived at ten features as the ideal feature set and reported the best performance (AUC: 0.80) by the support vector machines model. The authors also incorporated the model into an open-access

web application for individualized predictions and explanations. Similarly, Shi et al. (2012) used pre-operative PROMs to predict the long-term quality of life of patients after breast cancer surgery using two machine learning algorithms—artificial neural networks and linear regression—and found the neural networks model performed better than the linear model .

Exploring the possibility of decision-support regarding the self-referral of patients with low back pain to primary care, Nijeweme-d'Hollosy et al. (2018) evaluated three tree-based machine learning methods—decision tree, random forest, and boosted tree. The authors used fictive cases consisting of baseline pain-related PROMs to build prediction models of referral advice for the patients and found boosted trees gave the best performance (71% accuracy). Rahman et al. (2018) used their pain self-management mobile application "Manage My Pain" to collect a total of 130 PROMs from their mobile app users and found that pain volatility reported by the users could be predicted with an average accuracy of 70% using Random Forest. While the performance is modest, the model does perform better than a well-estimated clinical guess (Kattan et al., 2013). In their follow-up work, the authors reported achieving similar prediction accuracy (68%) with only nine features (Rahman et al., 2019). Schiltz et al. (2020) reported similar modest performance using random forest (AUC: 0.61) for predicting the risk of 30-day hospital readmission in more than six thousand older adults based on self-reports of activities of daily living limitations, co-morbidities in addition to patient demographics, socioeconomic and behavioural factors. The authors used the random forest algorithm to rank features in terms of their ability as predictors of hospital readmission and found activities of daily living limitations to be the single most valuable predictor. The study highlights how routine assessment of patient-reported data using machine learning methods can help identify patients at a higher risk of readmission.

Other clinical fields such as psychological disorders and dental care have also found machine learning methods valuable in their research. Andrews et al. (2017) evaluated the NANA toolkit they presented in their earlier work to determine whether or not the future depression status of the users, primarily older adults, can be predicted (Brown et al., 2018). The authors used six self-reported mood scores collected from the users with the help of the toolkit as features in a logistic regression (LASSO) model and achieved good predictive ability (AUC: 0.88) with only two predictors—*sadness* and *tiredness*. Kessler et al. (2016) reported modest accuracy in predicting long-term (10-12 years) depressive disorders in patients based on self-reported

depression-related measures using regression trees. The authors also found that machine learning models performed better with fewer predictors (between 9 and 13) than the traditional methods (23 predictors). Wang et al. (2020) explored two machine learning algorithms—extreme gradient boosting and naive bayes—to predict oral health status outcomes of young children based on self-reports and socio-demographics and reported modest prediction performance during validation.

Modern machine learning approaches offer several benefits over traditional methods. However, for high-stake applications, it is recommended that the models are externally validated to thoroughly evaluate their generalizability before being integrated into everyday clinical decision-making (Riley et al., 2016). Few studies in the existing literature have externally validated their machine learning implementations. This problem is often attributed to the lack of suitable datasets, non-data sharing, and ethical and legal constraints. Nijeweme-d'Hollosy et al. (2018) addressed the topic of generalisability by externally validating their prediction models, trained originally on fictive cases, using real-life cases, and found modest performance during testing and validation (71% and 72%, respectively). Chekroud et al. (2016) validated their machine learning prediction model using another clinical trial cohort. The gradient boosting model predicts whether or not a patient achieves clinical remission using anti-depressants based on baseline PROMs and gave a 64.6% accuracy in internal validation and 59.6% in the external validation. Considering that model accuracy often decreases during external validation, the model's performance is modest. In their follow-up work, the authors employed unsupervised learning in addition to supervised learning on the same two clinical cohorts as before to predict cluster-specific outcomes of anti-depressant treatment based on the baseline PROMs (Chekroud et al., 2017). They first clustered the patient profiles based on their baseline symptoms using a data-driven approach to identify three symptom clusters and then predicted cluster-specific short-term (10-12 weeks) treatment outcomes for patients. Upon external validation, the authors reported statistically above chance and clinically modest predictions in all the three cluster models.

The application of CBR methodology in the medical domain has a long history (Holt et al., 2005; Bichindaritz and Montani, 2009; Begum et al., 2010). Recent developments in CBR application have forayed into personalised physical activity plans (Smyth and Cunningham, 2017; Smyth, 2019). Smyth and Cunningham (2017) proposed a case-based approach to assist

marathon runners achieve their best race time. In the presented approach, the authors used the past race time of the runner and the race histories of other similar marathon runners to build a tailored race plan and predict their new personal best race time. They evaluated their approach using six years of race data from the London Marathon and achieved favourable results indicating that such a case-based approach can be of value for predicting and recommending activity routines. A sufficient description of the problem is also necessary for building robust CBR systems. Knowledge about the significance of various features in a dataset plays a critical role in building CBR systems. Unlike many other traditional machine learning methods such as ensemble and linear, CBR does not have implicit feature selection (Aamodt and Plaza, 1994; Weber et al., 2005). For clinical datasets that typically consist of many features and noise, building a CBR system elicits a reduction in the dimensionality of the dataset for non-redundant yet sufficient problem description and a reduction in computational costs. To that end, a great deal of work has been done to assess approaches for feature selection for building fuzzy rule-based CBR systems. Xiong and Funk (2006) presented a framework for selecting features based on the performance of fuzzy-rule-based CBR models. The authors later proposed a hierarchical approach where they used individual cases to optimise the possibility distributions in the case base and selected features based on the magnitude of their parameters in the similarity models (Xiong and Funk, 2010). A hybrid *wrapper-filter* approach was presented by Li et al. (2009) where the authors iteratively build CBR systems using feature subsets selected based on mutual information as a preset criterion and evaluated the implemented systems using a metric to determine the best set of features. In their previous work, the authors had used rough sets for feature reduction in large datasets (Li et al., 2006). Similarly, Zhu et al. (2015) determined reduced feature sets through neighborhood rough-set algorithm, which was applied in other similar published work for feature and case selection in CBR systems (Salamó and Golobardes, 2001; Salamo and Lopez-Sanchez, 2011).

2.4 myCBR

myCBR is an open-source tool for similarity-based retrievals, developed in a joint effort between the Competence Centre CBR at German Research Center for Artificial Intelligence (DFKI), Germany, and the School of Comput-

ing and Technology at the University of West London, UK (Stahl and Roth-Berghofer, 2008). *myCBR* includes a workbench that provides a powerful graphical user interface (GUI) for developing rapid prototypes of similarity-based retrieval systems and a software development kit (SDK) to easily integrate the prototypes into independent applications (Bach and Althoff, 2012). Bach et al. (2019) gave a demonstration of *myCBR SDK* functionality and how it can be used for rapid development of CBR applications.

myCBR workbench has a *case bases* view and a *modelling* view. Figure 2.1 and 2.2 present the *case bases* and the *modelling* view of the *myCBR workbench*. *myCBR workbench* supports importing cases from a csv file. In the *case bases* view, the available case bases, stored cases, and the individual case structure can be viewed. Figure 2.1 shows a case base named *CB_csvImport* under the concept *Participant* which was populated by importing cases from a csv file. Each case is represented by its attributes, as shown by the example *Participant0* which has six attributes of data type *float*. Case structure and similarity measures can be created in the *modelling* view of the *workbench*. Local similarity measures for each individual attribute can be created and defined in the *workbench*, as shown by the example attribute *cycling* with a corresponding local similarity measure *cyclingSim* in figure 2.2. Once the local similarity measures for all the attributes have been created, the global similarity measures can be defined. Figure 2.3 shows the global similarity definition for the concept *Participant*, a global similarity measure named *ParticipantSim* with equally weighted local similarity measures.

myCBR has previously been used to build knowledge models for cooking recipe recommendations (Bach et al., 2012), customer service support for machine diagnosis (Bach et al., 2011), audio advisor (Sauer et al., 2013b) and workflow recommendations for gold ore treatment (Sauer et al., 2013a) among others. For the research in this thesis, we used *myCBR workbench* to build CBR models for two healthcare datasets and *myCBR SDK*[1] to integrate the models into the *JupyterLab*[2] python environment to carry out our experiments.

[1] $www.github.com/ntnu-ai-lab/mycbr-sdk$
[2] $www.jupyter.org$

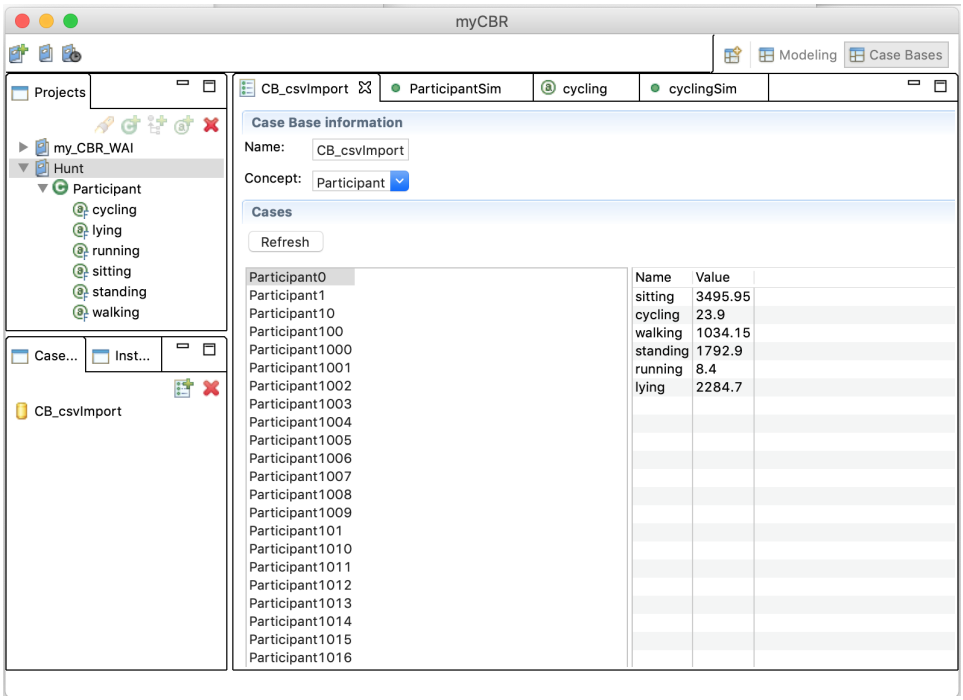


Figure 2.1: The *Case bases* view of the *myCBR workbench* shows the open projects, the case structure for concept *Participant* under a sample project named *Hunt* on the top left pane, available case bases in the pane below, and case instances on the right.

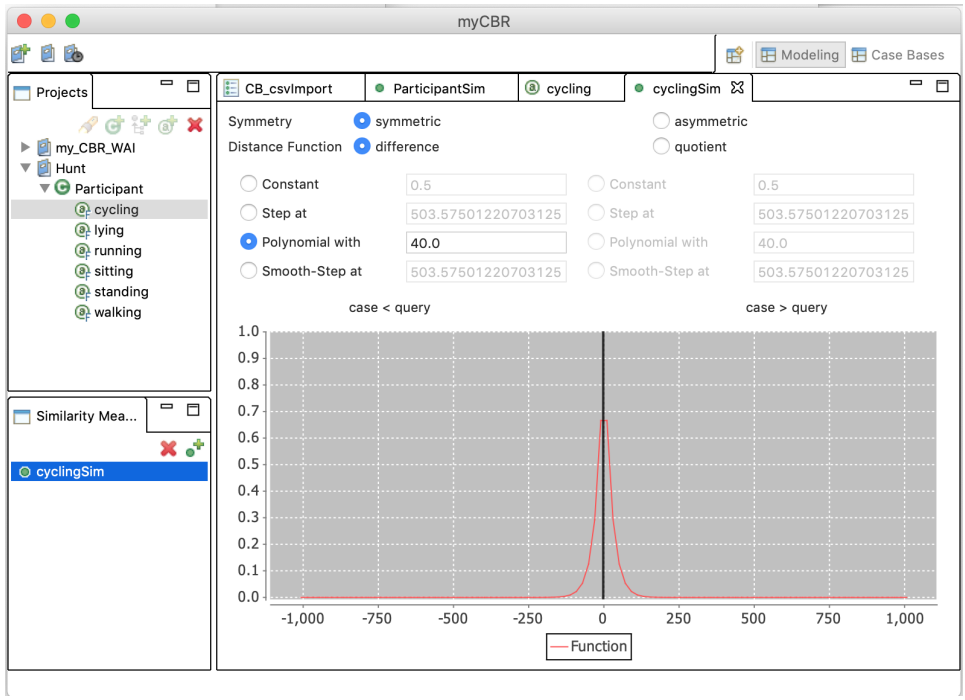


Figure 2.2: The *Modelling* view of the *myCBR workbench* showing the available similarity measures for the selected attribute on the bottom left pane and the definition of the selected similarity measure on the right.

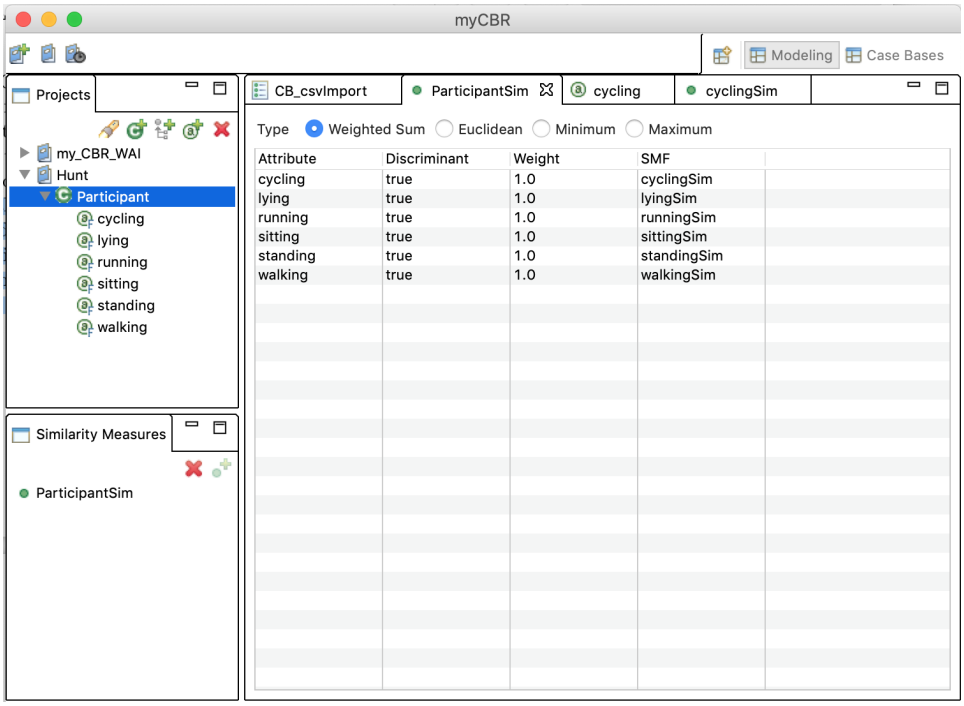


Figure 2.3: The global similarity definitions in the *modelling* view of *myCBR workbench*.

Chapter 3

Corpora

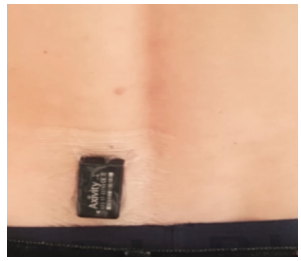
The research in this doctoral work focused on facilitating further development of CDSS using data from population-based and intervention-based studies by exploring the capabilities of predictive analytics using machine learning methods and the evaluation thereof. We worked with two datasets: (1) HUNT4 cohort data (objective physical behaviour data of the HUNT4 study) and (2) SELFBACK user data. Using these two datasets, we attempted to drive our research towards exploring ways in which value can be derived from these for future CDSS.

3.1 HUNT4

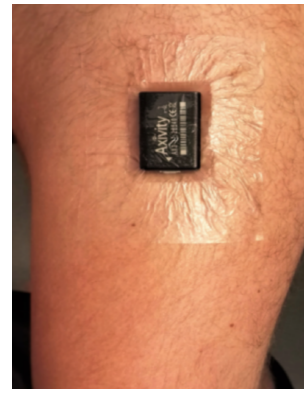
The first dataset used in this doctoral research work is the objectively measured physical activity data collected during the HUNT4 [1] cohort study. The HUNT study is carried out in mid-Norway and is one of the largest cohort studies of its kind. The previous three studies (HUNT1 1984-86, HUNT2 1995-97 and HUNT3 2006-08) collected health data, mainly through questionnaires and clinical examinations which has been used extensively for further epidemiological research. For the first time, HUNT4 also collected objective physical activity data through body-worn accelerometers.

[1] *www.ntnu.no*/HUNT4

Figure 3.1: The Axivity AX3 accelerometer



(a) Lower Back Placement



(b) Upper Thigh Placement

Figure 3.2: Placement of the accelerometers on the participants

3.1.1 Equipment and Setup

The sensor used for collecting physical activity data was the AX3 Axivity[2] accelerometer, as shown in figure 3.1. The device is a tri-axial accelerometer weighing 11 grams with dimensions of 23 x 32.5 x 7.6 mm. The data was recorded at a sampling frequency of 50Hz. Two of these accelerometers were placed on every participant, one on the lower back and another on the upper thigh, as shown in image 3.2. More information about the data collection setup can be found in Reinsve and Bach (2018).

[2] *www.axivity.com/product/ax3*

3.1.2 Data Collection

All residents in the Nord-Trøndelag county aged 13 year or above were invited to participate in the study. Those who volunteered to participate in the objective physical activity data collection were fitted with two tri-axial accelerometers and wore them for a period of seven consecutive days. The sensors record the vibrations, movement and orientation changes in the three axes. Figure 3.3 presents a standard human activity recognition pipeline using sensor data. The raw data is downloaded from the accelerometers and classified into 17 different physical activities using two pre-trained machine learning models- Support Vector Machines for synchronizing data from the two sensors and Random Forest classifier to classify the activities (Vågeskar, 2017; Bach et al., 2021). The resulting data set contains the H4ID (unique ID for each HUNT4 participant), number of minutes of each different activity, the date and day of the week in a csv file. The physical activities are later merged into six main categories, presented in table 3.1. The data collection in HUNT4 spanned over 18 months and was concluded in February 2019. As a result, objective measurements of a total of 35449 participants have been collected and basic physical activities have been assigned.

Table 3.1: Activity Descriptions.

Activity	**Description**
Lying	The person is lying down
Sitting	When the person's buttocks is on the seat of a chair or something similar
Standing	Upright, feet supporting the person's body weight
Walking	Locomotion towards a destination with one or more strides
Running	Locomotion towards a destination, with at least two steps where both feet leave the ground during each stride
Cycling	The person is riding a bicycle

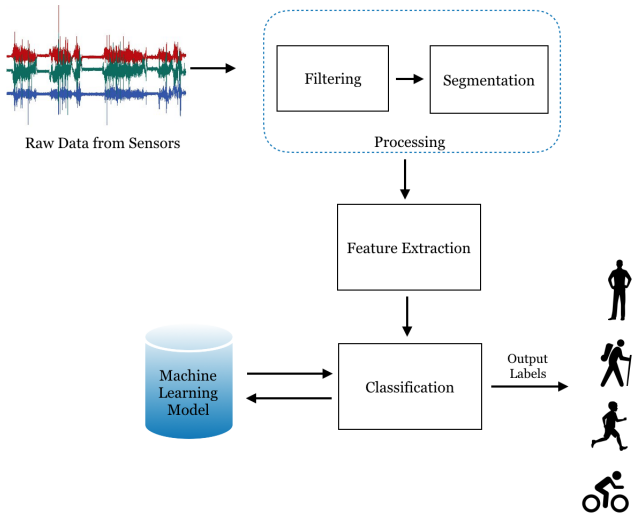


Figure 3.3: Human Activity Recognition process using raw data from tri-axial accelerometers.

3.1.3 Data Pre-processing

Personal identifiers of the participants are transformed using a hash operator and the original identifiers are removed in order to de-identify the data. To prepare the dataset for further analysis, it is processed such that equal amount of data is present across the participants. This is due to several reasons. Firstly, even though all the participants were expected to wear the sensors for seven consecutive days, some had to either remove the device due skin irritation or the sensor malfunctioned, leaving the number of days of collected data less than seven for some participants. Therefore, the dataset was processed to obtain the same amount of data for each included participant and the decision was the number of days. It was decided to include only the participants who have full six days of measured data, since this number was much higher than the total number of participants with full seven days of measured data. Further, to eliminate any classifications errors or sensor malfunctions, any records containing zero minutes for *lying, standing, sitting* and less than one minute for *walking* activity as well as records where the sum of all activities exceeds 1440 minutes for a day (which represents the total minutes in a day) were removed. Following the data processing, the

number of participants in the dataset came down to 31,113 out of the 35,449 originally.

Figure 3.4 presents the distribution of the different physical activities in the final dataset. The first boxplot shows the regular quartile distribution and it becomes immediately evident that the activities *running* and *cycling* are in much smaller quantity in the dataset and therefore are not visible on the given scale. To invert this effect, log operator is applied before plotting the dataset and subsequently, the second boxplot shows log distribution of the physical activities.

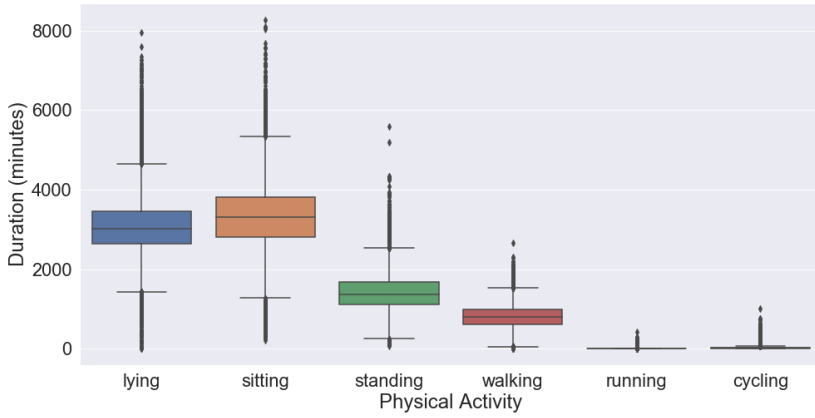
3.2 SELFBACK

The SELFBACK [3] project was funded by the European Union Horizon 2020 research and innovation program (under grant agreement no. 689043) and began in January 2016. The overall aim of the SELFBACK project is to provide digital interventions for self-management of pain-related symptoms to patients with non-specific musculoskeletal disorders, specifically low back pain (LBP). The project involves development, implementation and further evaluation of effectiveness of the SELFBACK system, a mobile decision support system (DSS) application that can provide tailored self-management plans for the users of the mobile application. The weekly self-management plans are tailored based on questionnaires answered by the user at certain time-points and include components of recommended amount of physical activity (daily number of steps), strength and flexibility exercises as well as education to motivate the users. The daily number of steps are recorded by a wearable device while completion of the recommended exercises and educational readings are self-reported in the mobile application. Data collected during two randomized controlled trials (RCT) has been used in this doctoral research (Sandal et al., 2019; Marcuzzi et al., 2021).

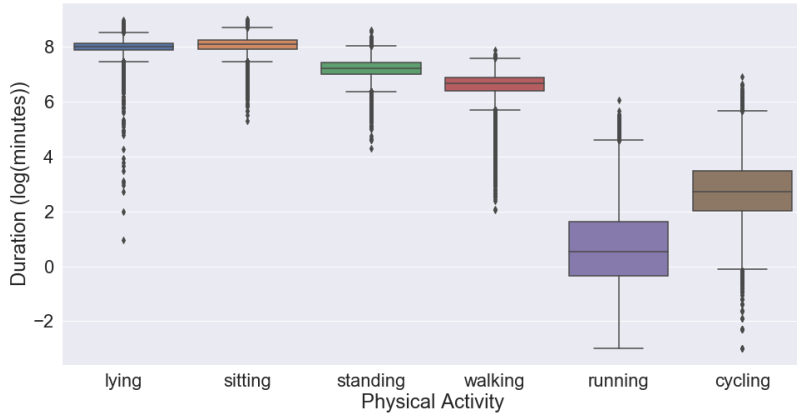
3.2.1 Equipment and Setup

The SELFBACK project involved collection of both subjective and objective data from the recruited participants. Physical activity measurements were recorded using a wearable activity tracker (Mi Band 3, Xiaomi), as shown in figure 3.5. The device recorded the number of steps achieved per day.

[3] *www.selfback.eu*



(a) Regular Distribution



(b) Log Distribution

Figure 3.4: Boxplots presenting the distribution of physical activity data of the participants in the HUNT4 dataset based on the quartiles.

Figure 3.5: Xiaomi Mi Band 3 used in the SELFBACK project for collecting daily step count

Patients also filled several questionnaires, first at the time of recruitment (web-based questionnaire) and then throughout using the SELFBACK mobile application.

3.2.2 Data Collection

Figure 3.6 gives an overview for the data collection efforts in both the trials.

(a) RCT I
This single-blinded, two-armed RCT[4] was aimed at testing the effectiveness of the SELFBACK DSS with usual care (intervention) against usual care only (control group) for patients with non-specific LBP. Care-seeking patients with non-specific LBP were recruited through the referral of their primary care clinician (i.e., physiotherapists, chiropractors, general practitioners) in Trondheim, Norway and Odense, Denmark. Patients were screened for eligibility based on a preset inclusion/exclusion criteria that can be found in Sandal et al. (2019). The eligible patients were invited to participate in the RCT and those who accepted the invite answered a baseline questionnaire. The participating patients were randomized into either intervention group or control group. The intervention group had access to the SELFBACK DSS mobile application and received tailored self-management plans weekly whereas the control group did not. The participants answered follow-up questionnaires at 6-weeks, 3 months, 6 months and 9 months, in addition to the baseline and weekly tailoring questionnaires (only in the intervention group). A total of 461 participants were included in this trial.

[4] *www.clinicaltrials.gov/ct2/show/NCT*03798288

(b) RCT II
This is a single-blinded, three-armed RCT[5] aimed at evaluating the effectiveness of the SELFBACK tailored self-management interventions against a web-based self-management intervention without tailoring or usual care in people with low back and/or neck pain. The recruitment was carried out in the multidisciplinary outpatient clinic for back, neck and shoulder rehabilitation at the St. Olavs Hospital in Trondheim, Norway. Referred patients that were accepted for treatment at the clinic were screened for eligibility based on a preset eligibility criteria that can be found in Marcuzzi et al. (2021). The eligible patients were invited to the study and those who accepted were randomly assigned into one of the three arms with equal allocations: 1) SELFBACK app with usual care, 2) web-based intervention with usual care, and 3) usual care only. A total of 294 participants were included in this trial. Self-reported outcomes were collected through web-based questionnaires at three follow-ups- 6 weeks, 3 months and 6 months, in addition to the weekly tailoring questions for those who had access to the SELFBACK mobile app.

The questionnaires consist of various different self-reported measures of *pain intensity, pain self-efficacy, physical activity, functional ability, work-ability, sleep quality, fear avoidance* and *mood*. Additionally, the baseline questionnaire also included patient sociodemographics such as age, height, weight, gender, education, employment, living situation and family. Tables 3.2 and 3.3 summarise the information collected from the participants at various time points in the clinical trials.

3.2.3 Data Pre-processing

Before using the patient-reported data for further analysis, personal identifiers assigned to each participant are removed to anonymize the data. For the scope of this doctoral work, only the baseline data and follow-up data was used in the experimental analysis. The data is further processed to identify and remove records with empty value in some or all fields, no or missing baseline data, no or missing follow-up data. Following data processing, the RCT I dataset consisted of PROMs from 376 participants (218 in intervention, 158 in control) while RCT II consisted of PROMs from 75 participants (only the SELFBACK app with usual care group).

[5] *www.clinicaltrials.gov/ct2/show/NCT*04463043

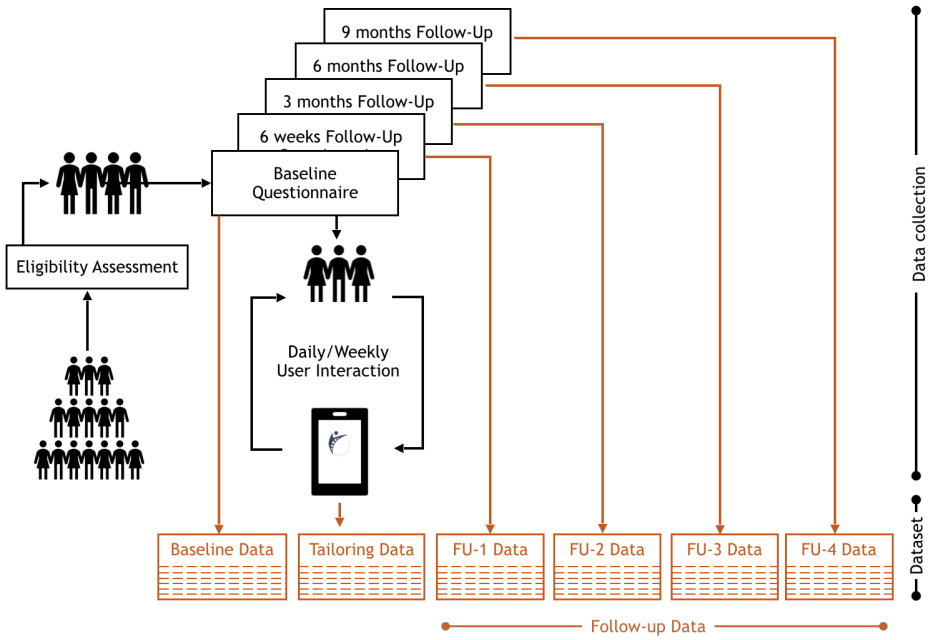


Figure 3.6: Overview of data collection in the SELFBACK RCTs. The different data components are indicated by the orange boxes.

Table 3.2: Various Patient-Reported Outcome Measures in RCT I

Descriptive variables		
Patient Characteristics	Sociodemographics	
Primary Outcome Measure		
Roland Morris Disability Questionnaire		
Secondary Outcome Measures		
Pain Self-Efficacy Questionnaire	Fear Avoidance Belief Questionnaire	Pain Intensity
Brief Illness Perception Questionnaire	Saltin-Grimby Physical Activity Level Scale	
Global Perceived Effect		
Other Outcome Measures		
Workability	Health-related Quality of Life	Activity Limitation
Patient Health Questionnaire	Perceived Stress Scale	Sleep
Patient Specific Functional Scale	Pain Duration and frequency	Physical Activity
Exercise Volume	Virtual Care Climate Questionnaire	User ratings

Table 3.3: Various Patient-Reported Outcome Measures in RCT II

Descriptive variables		
Patient Characteristics	Sociodemographics	
Primary Outcome Measure		
Musculoskeletal Health Questionnaire		
Secondary Outcome Measures		
Roland Morris Disability Questionnaire	Neck Disability Index	Pain Intensity
Health-related Quality of Life	Pain Self-Efficacy Questionnaire	
Brief Illness Perception Questionnaire		
Other Outcome Measures		
Fear-Avoidance Belief Questionnaire	Perceived Stress Scale	Sleep Problems
Patient Health Questionnaire-2	Patient Specific Functional Scale	
Saltin-Grimby Physical Activity Level Scale	Global Perceived Effect	Work ability index
Patient Acceptable Symptom State	Health care consumption	Sickness absence

Chapter 4

Methodology

Designing targeted interventions or personalised activity plans requires considerable clinical expertise and evidence-based research to support clinical decision-making. Patient-centred CDSS has the potential to support clinical practice. Some of the most common and important questions a patient often has are—*"What is the cause of my health problems?"*, *"What can I do to get more active to achieve better health?"*, *"Am I going to get better?"*, *"When will I get better?"* —among others. While there are no black and white answers to these questions, evidence-based research and patient-centred CDSS can support the healthcare provider in keeping the patient in the loop (Harrell Jr et al., 1996).

In the research leading up to this thesis, we undertook a data-driven approach to address the matter of developing methods for patient-centred systems that can utilise intervention-based and population-based healthcare datasets to facilitate the clinical decision-making process. With the application of data-driven methodologies and access to expert knowledge, we endeavoured to transform the information stored in intervention-based and population-based healthcare datasets into case bases and develop case-based models that can add value to the development of future patient-centred and predictive CDSS. From developing case representations and similarities to exploring different feature selection methodologies, we incorporated domain expert knowledge into the steps involved in building a case-based model that can efficiently utilise a healthcare dataset. We also explored various conventional machine learning methods in addition to CBR to better comprehend the differences in the working and suitability of the two approaches for our problem domain. In this chapter, we describe the population of the case base,

and the representation of each case in the case base in *myCBR workbench*, the methodology for developing the local and global similarities, the feature selection methodologies applied to the datasets presented in the last chapter, and finally, the models developed. Figure 4.1 presents an overview of the approach undertaken in developing learning models for better utilising healthcare datasets in this work.

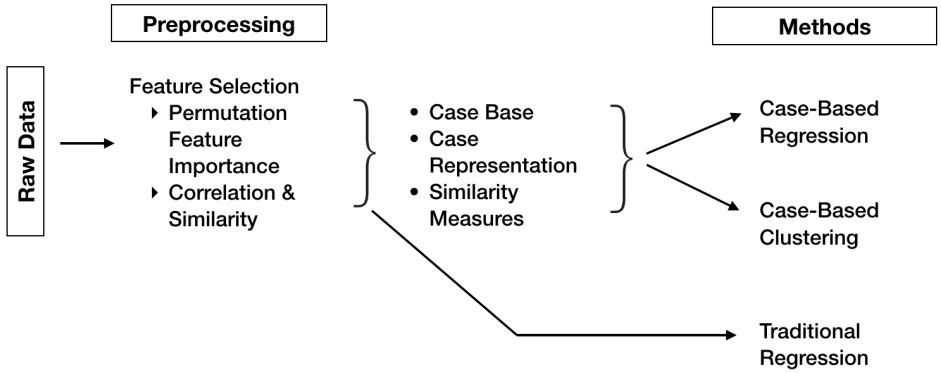


Figure 4.1: An overview of the process of developing learning models using machine learning methods applied to the datasets in this thesis.

In the sections that follow, we describe the case base, case representation, and the similarity measure development first and then the feature selection methods for two reasons. First, the main focus of this doctoral research is the development of CBR systems for healthcare datasets, and therefore, it is logical to explain the said part first. Second, the similarity measures play a significant role in the feature selection section.

4.1 Case Base and Case Representation

The case base is an essential component of any CBR system as the case base forms the basis for any future problem-solving (Aamodt and Plaza, 1994; Richter and Weber, 2013). The case base is, as the name suggests, a collection of cases storing previously solved problems and their solutions. The case base is organized to facilitate retrieval of the most similar cases in the event of arrival of a new problem. When a new problem arrives, the system

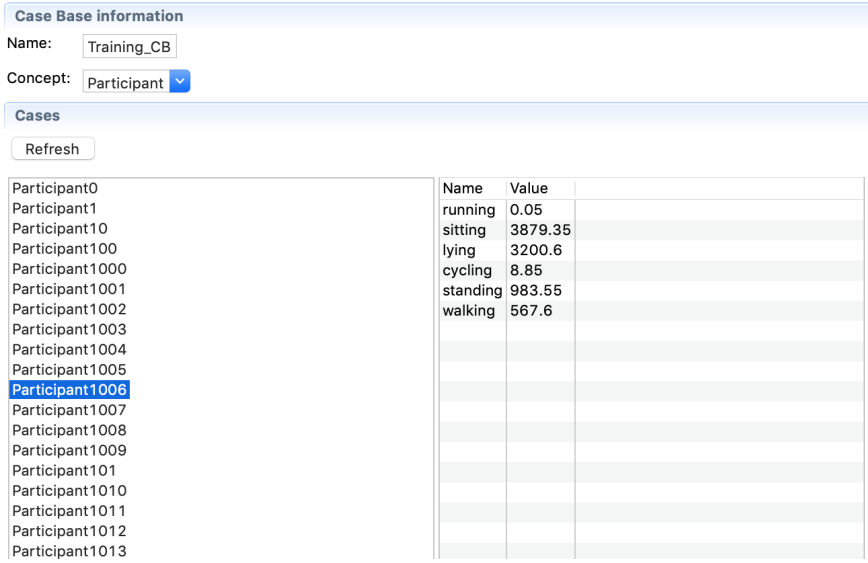
searches its case base to look for similar past cases. The solution of the retrieved past case(s) provide a starting point for generating a solution for the new problem. A quality case base is critical for the success of a CBR system since without the prior problem-solving experiences (cases), the system becomes vain (Richter and Weber, 2013). The information stored in each individual case also impacts the performance of the system and to ensure that only the most valuable information goes into each case and into the case base as a whole, data pre-processing and feature selection are necessary steps.

As cases form the basis of a CBR system, representation of the cases is an integral and important part of developing CBR systems. Case representation here refers to the way the data is represented in the CBR system, and as any form of representing data can be considered as case representation, it comes in many different forms. The intended functionality of the system and the ease with which the information can be acquired for representing the cases are two important measures that need to be considered while deciding the information to be stored in the cases and a suitable case representation. The simplest and the most commonly used way is by using *feature-value* pairs (Richter and Weber, 2013).

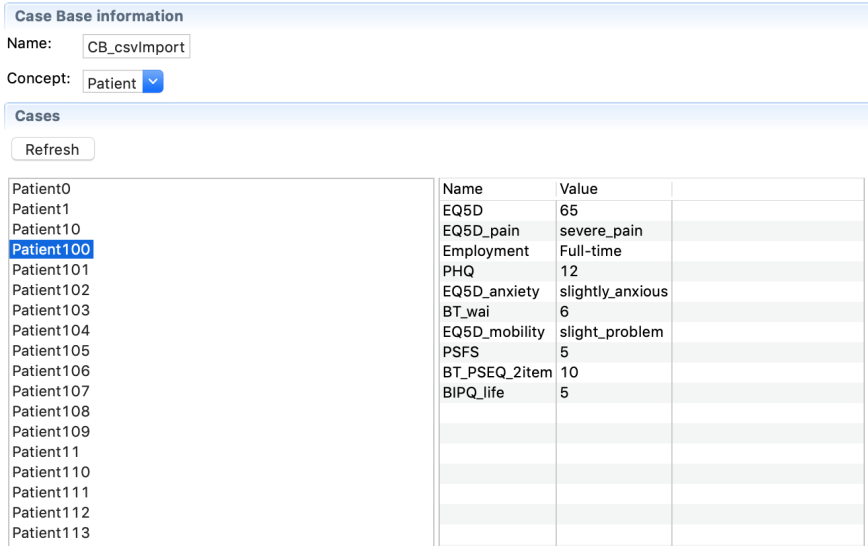
Figure 4.2 presents examples of case base and case representation in *myCBR workbench* for datasets SELFBACK RCT I and HUNT4 , respectively. The *name* field at the top left corner of each figure shows the name of the case base, and below it, the name of the *concept*. Under the *concept* field is a list of the individual cases stored in the case base. The right part of each figure shows an individual case chosen at random. The *name* column shows the name of each feature while *value* displays the value contained by the respective feature.

4.2 Similarity Measure Development

Similarity measures, as discussed in the previous chapter (section 2.1), play an important role in determination and retrieval of similar cases from the case base. Therefore, considerable emphasis is put on the development of suitable similarity measures in our research work. Modelling the similarity measures for any specialised application domain can be a challenging task. Firstly, the system developers have to balance the input from the domain experts and the available data. And secondly, they have to identify important attributes to be included in the knowledge model to avoid includ-



(a)



(b)

Figure 4.2: Examples of populated case base on the left and case representation on the right for CBR systems for the **a.** HUNT4 and **b.** SELFBACK datasets.

ing redundant information. The entire process requires close collaboration with the domain experts. Having a criteria that can lead the knowledge modelling process can ease the burden and be helpful for both the parties. Using a data-driven approach can stratify this process. Keeping in mind the *local-global principle* for similarity modelling (Richter and Weber, 2013), we develop the local similarities first, followed by the global level similarities.

The similarity measures developed must capture the domain knowledge and approximate a *utility function* such that it estimates the utility of the cases and finds suitable cases from the case base for problem-solving. If each case is represented using a numerical and a categorical attribute, the assignment of similarity behaviour would be different for either attribute to reflect the implicit knowledge stored in the attribute's behaviour. A categorical attribute, for instance, may have an implicit order in the values it can take in each case and so, the order must be captured and preserved by the similarity measure. On the other hand, setting the upper and lower limits for a numerical attribute may be straightforward, assigning the similarity behaviour, such that it captures the data distribution of the attribute, is not.

4.2.1 Local Similarity

We discuss the process of similarity modelling for the numerical attributes first, followed by the categorical ones (the order is not relevant to the modelling process), and assume that numerical local similarity measures are polynomial distance functions. The goal then is to determine the degree of the polynomial function such that it covers the entire similarity range [0,1] while capturing the similarity behaviour of the attribute. Creating a box-plot of the dataset that shows the data distribution of each individual (continuous) numerical attribute allows modelling the similarity measure of each individual attribute, as we can use the Inter Quartile Range (IQR) and the range (min to max) for transferring the knowledge into modelling their similarity behaviour. From the box plot, the quartiles Q_1 and Q_3, which indicate the majority spread for the dataset, can be used as reference points to set how steep the decline in the polynomial function, or alternatively, the decrease in the similarity should be. IQR represents the difference between upper (Q_3) and lower (Q_1) quartiles in the box-plot, that is $IQR = Q_3 - Q_1$. Taking the example of HUNT4 dataset, figure 4.3 shows a box-plot at the top presenting the distribution of all the attributes in the dataset. Eq. 4.1 describes how the reference points r_1 and r_2 relate to the IQR and range

(determined using IQR*1.5 method) for each individual attribute.

$$\begin{aligned} r_1 &= IQR \\ r_2 &= range \end{aligned} \tag{4.1}$$

Since the similarity functions are assumed to be symmetrical, the polynomial degree of the similarity function y (red line on the graph in figure 4.3) can be adjusted such that

$$\begin{aligned} y(r_1) &\approx 0.30 \\ y(r_2) &\approx 0 \end{aligned} \tag{4.2}$$

The bottom part of figure 4.3 shows how the similarity function varies after applying the methodology in equation 4.1 and 4.2. The arrows present how the quartiles for *sitting* relate to the decrease of similarity at a certain distance. The bigger the polynomial degree, the steeper the similarity function and the more precise the attribute values in the retrieved cases. The decline in the similarity function is steeper in the beginning until at r_1 it reaches close to $y(r_1)$ and then decreases gradually until at r_2 it is approximately close to $y(r_2)$. This step ensures that the similarity function covers the entire attribute range and the similarity measure range $[0,1]$. The choice of $y(r_1)$ and $y(r_2)$ depends on the domain expert's knowledge and satisfaction with the effect. We, however, experimented with different values and found these best suited for our application domain. Furthermore, the symmetrical function ensures that the similarity measures can later be used as a metric for clustering the case base.

While the local similarity measures for numerical attributes can be derived using their data distributions, assigning the similarity behaviour for categorical attributes can be challenging since it depends on whether or not there is an existing relationship between the values. The SELFBACK datasets, for example, comprise categorical attributes and need to be handled differently from the numerical ones. The local similarity measures for categorical attributes must model and preserve the relationship amongst the values (for ordinal attributes) while achieving the desired variation in the similarity measure in the range $[0,1]$, as shown in figure 4.4. The ordered symbolic function, shown in figure 4.4a, ensures that the order between the values is preserved, and the polynomial function ensures that the entire similarity range is covered. In figure 4.4b, seeing as there is no specific order or

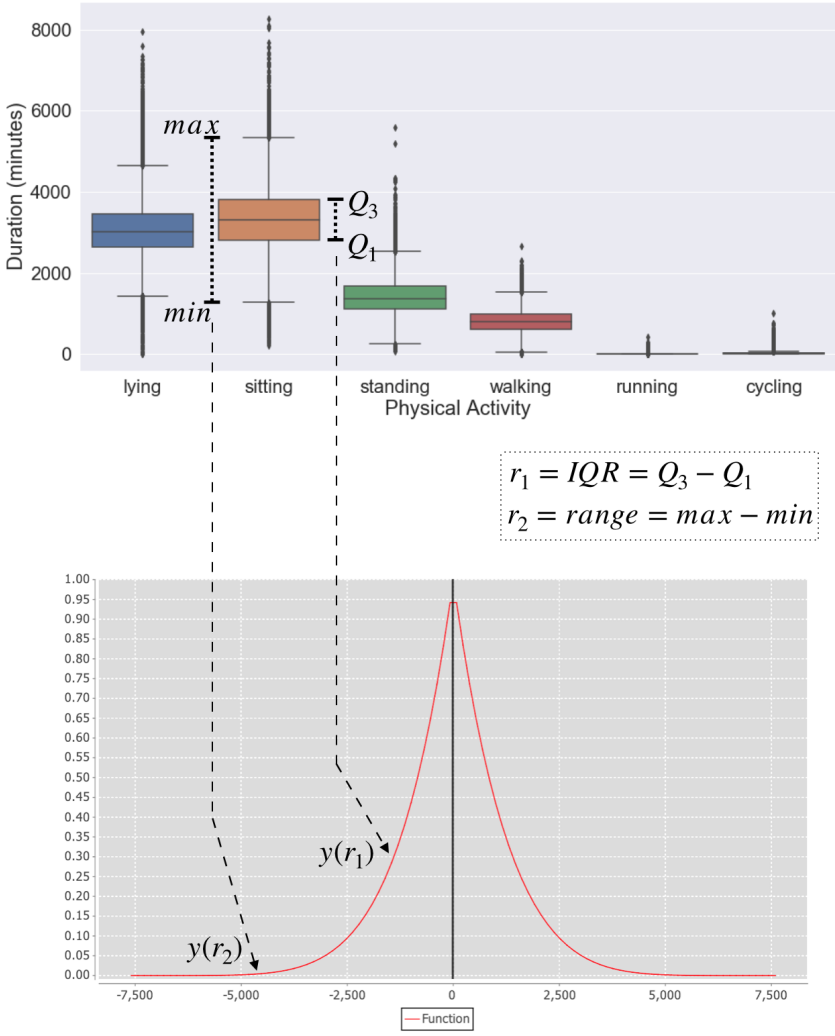
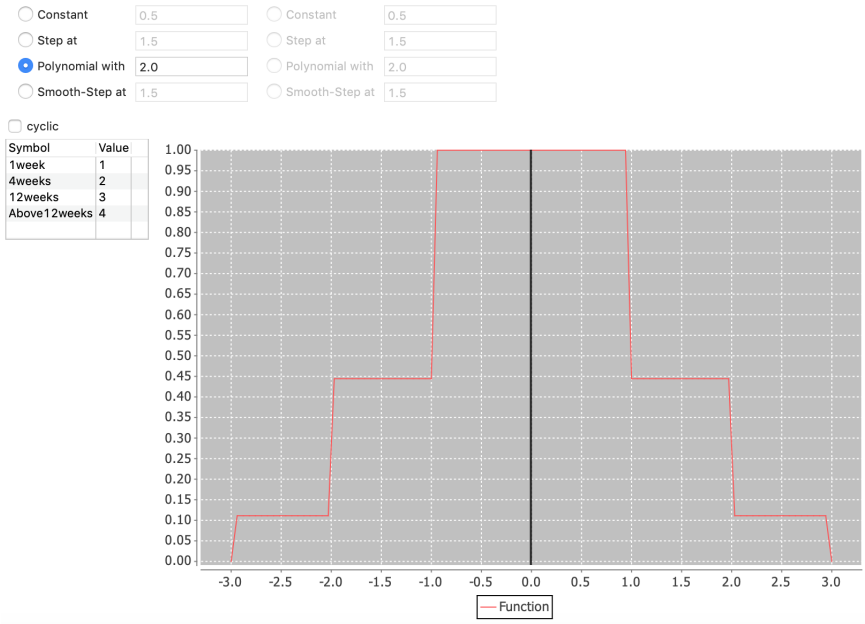


Figure 4.3: Example for data-driven local similarity modelling: On the bottom is a screenshot of a polynomial similarity function for the attribute *sitting*.

relation in the attribute values, the similarity of one value to every different value has been set to zero.



(a)

	Full-time	Part-time
Full-time	1.0	0.0
Part-time	0.0	1.0

(b)

Figure 4.4: Example of local similarity modelling for categorical attributes in the SELFBACK dataset. **a**. Polynomial similarity function for ordinal attribute *Pain_current*. **b**. Symbolic Similarity function for nominal attribute *Employment*.

4.2.2 Global Similarity

The attribute-level similarity measures reflect only the local view. A comparison of the entire concept, or participant in our datasets, requires a global view. Using the weighted sum approach, the global similarity measure *globalSim* for an object A with a set of n attributes a can be formulated as follows:

$$globalSim(A) = \sum_{i=1}^{n} \omega_i localSim(a_i) \tag{4.3}$$

where $localSim(a_i)$ is the local similarity measure for attribute a_i and the weight ω_i represents the influence the attribute a_i has over the global similarity measure. The weights determine how much each attribute contributes to the global view. Figure 4.5 shows an example of global similarity measures where every individual local similarity measure is weighed equally for the HUNT4 and SELFBACK datasets.

Type ◉ Weighted Sum ○ Euclidean ○ Minimum ○ Maximum

Attribute	Discriminant	Weight	SMF
cycling	true	1.0	cycling_sim
lying	true	1.0	lying_sim
running	true	1.0	running_sim
sitting	true	1.0	sitting_sim
standing	true	1.0	standing_sim
walking	true	1.0	walking_sim

(a)

Type ◉ Weighted Sum ○ Euclidean ○ Minimum ○ Maximum

Attribute	Discriminant	Weight	SMF
BIPQ_life	true	1.0	bipq_l
BT_PSEQ_2item	true	1.0	btpseq
BT_wai	true	1.0	btwai
EQ5D	true	1.0	eq5d
EQ5D_anxiety	true	1.0	eq5d_an
EQ5D_mobility	true	1.0	eq5d_m
EQ5D_pain	true	1.0	eq5d_p
Employment	true	1.0	empl
PHQ	true	1.0	phq
PSFS	true	1.0	psfs

(b)

Figure 4.5: Example of global similarity measures for the datasets **a**. HUNT4 **b**. SELFBACK

4.3 Feature Selection

One of the major goals in application areas of machine learning often is to determine which features contribute the most to certain predictions. In the healthcare application domain, it becomes clear why there is a need for feature selection given that the data is vast and not all of it is necessary for

decision making (Colaco et al., 2019). From machine learning perspective, adding more information can also reduce the generalizability of the model(s) by increasing the overall complexity and lead to reduced performance and transparency (Jović et al., 2015). The existing methods for feature selection are often categorized into one of the following three based on the restrictions they impose on the machine learning methods: *filter, wrapper* and *embedded* (Kira and Rendell, 1992). Linear and tree-based learning methods are one of the most convenient and popular approaches of feature selection in the existing literature.

During the course of research in this doctoral work, we explored with different approaches for feature selection. Figure 4.6 presents two such strategies that were used in our work and found to be the most useful among others following evaluation. On the left side of the figure is the feature selection using embedded method feature importance using XGBoost regressor, while on the right side of the figure is the proposed hybrid method which combines correlation with data-driven similarities to drive the feature selection process.

4.3.1 Importance-based Feature Selection

Feature Importance is one of several ways of feature selection and refers to a calculated score indicating the relative importance of a feature in the performance of a machine learning model (Zien et al., 2009). Tree-based algorithms often compute feature importance scores based on gini impurity where the relative importance of the features is assessed based on their relative rank (i.e. depth) in the decision tree (Nembrini et al., 2018; Pedregosa et al., 2011). The features at the top of the decision tree contribute to the prediction of a larger fraction of the input samples and therefore, have a higher feature importance. Impurity-based feature selection is simple and fast to compute, but suffers from flaws. This method is heavily biased in favour of features with several possible split points and high cardinality, and can produce biased results in case of high correlation amongst the features, that is, in case of correlated features, the impurity-based method may select one feature and ignore the other entirely.

The drawbacks of impurity-based feature importance can be overcome by estimating the importance of each feature by shuffling its values and computing the impact this action has on the model's prediction performance (Altmann et al., 2010). This method is known as *Permutation feature impor-*

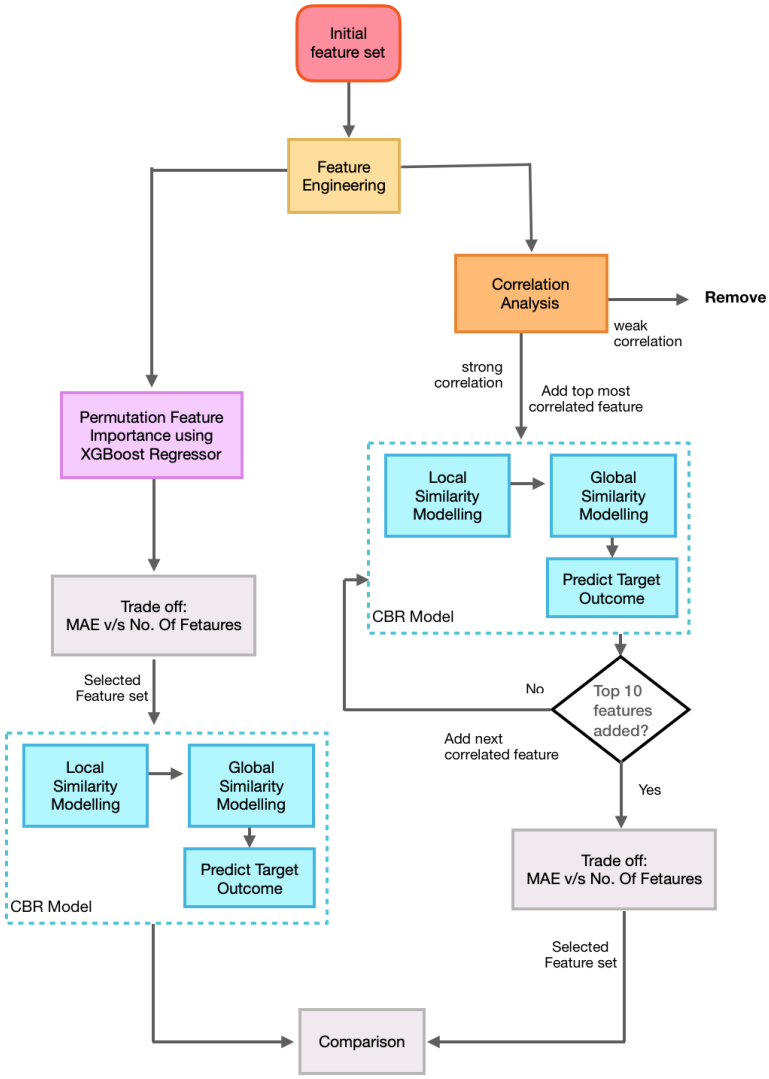


Figure 4.6: Feature Selection Process. **MAE**: Mean Absolute Error

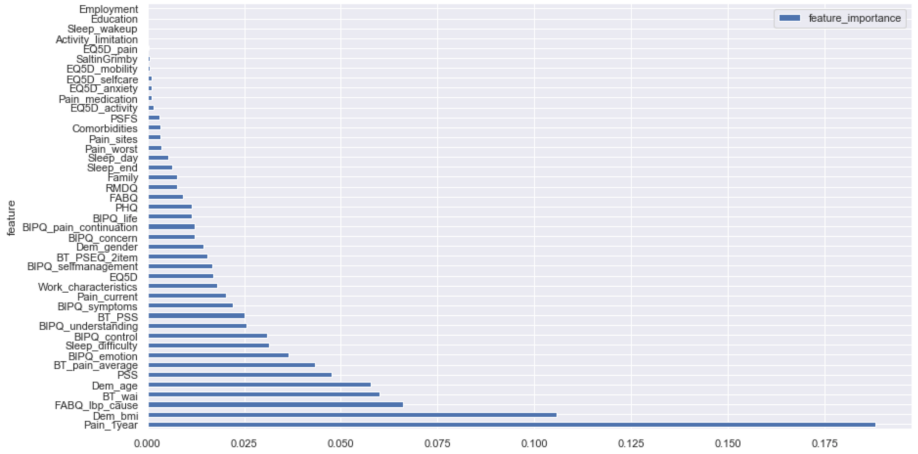
tance and is rather straightforward. First, a baseline model is trained with all the features of the dataset and the prediction performance of the model is recorded using an evaluation metric, say mean absolute error in prediction. Then, the values of one feature in the dataset are permuted (or in simpler terms, shuffled) and the model is re-trained and the error in prediction is estimated for the modified dataset. The process is repeated one feature at a time until all the features in the dataset have been permuted and the prediction error has been recorded. As it is, the most important features would be the ones that have the greatest impact on the model's prediction error when permuted.

As presented on the left part in figure 4.6, Permutation feature importance using XGBoost as the base regressor was used for estimating importance of various features in the SELFBACK dataset, and thereafter, selecting optimal feature sets based on the trade-off between mean absolute error in prediction and the number of features. Figure 4.7 presents an example of feature selection using the discussed approach for one of the target outcomes—*Global Perceived Effect (GPE)* (input: baseline data, target: GPE at follow-up 1)—from the SELFBACK RCT I dataset.

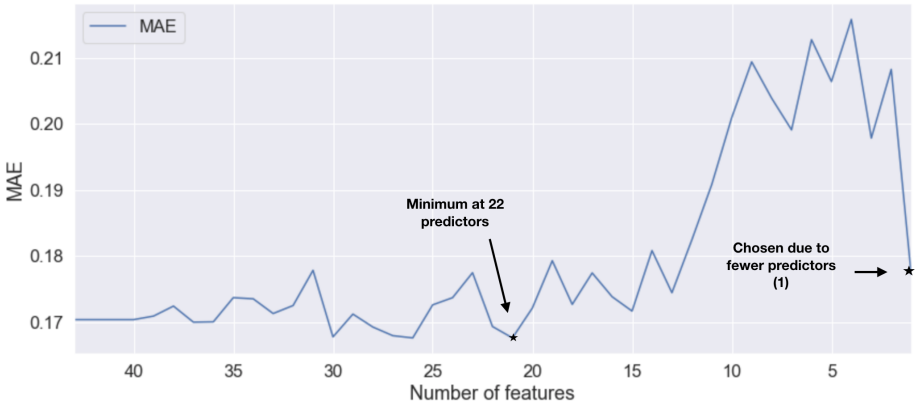
4.3.2 Correlation and Similarity-based Feature Selection

Correlation refers to the degree of linear relationship between two variables, either positive or negative. Take for example housing prices. Price of a house is determined by several factors and is positively correlated to the number of bedrooms, while is negatively correlated to the crime rate in the surrounding area (Raya et al., 2012). The more the number of bedrooms, the higher the price is expected to be, the higher the crime rate, the lower the price is expected to be. Correlation-based feature selection is a filter method approach and as it is, independent of the type of learning algorithm used for predicting the target.

To find a minimal feature subset for developing a CBR system, a prototype of the system can be used in the process, as this way, we know what we are optimizing against. Our interest lies in maximising the *utility* of the output. Therefore, in order to tailor the feature selection to the development of CBR systems, a hybrid approach that uses correlation and data-driven similarities can be applied to learn the minimal feature subset. Figure 4.6



(a)

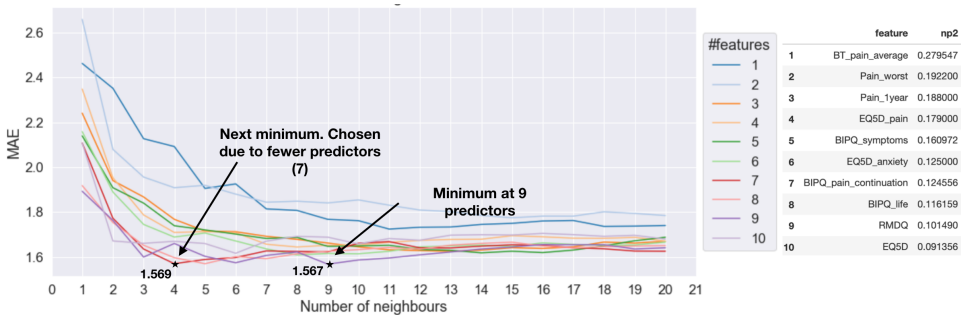


(b)

Figure 4.7: Importance-based Feature Selection. **a.** Features ranked by their importance. **b.** Effect of feature permutation on the base regressor XGB. The MAE (mean absolute error) on the y-axis in this plot is scaled to fit the range [0,1].

(the right side) shows the proposed feature selection pipeline using a hybrid *correlation-similarity*-based approach for selecting minimal feature subsets. First, correlation is estimated between all the features and the target outcome. Then based on their correlation coefficient and the computed p-value, the top correlated features (significant at the 0.05 level and limited in num-

ber to stratify the modelling process) can be filtered out for further development of the CBR system in the myCBR workbench. Using a data-driven similarity modelling approach (explained in section 4.2), local similarities can be modelled for each feature individually. The local similarities play an important role in capturing the value range of each individual feature and subsequently, tailoring the similarity behaviour of the features included in the CBR system. Therefore, attention is paid to the development of the local similarity measures for each individual attribute before evaluating the entire system. Figure 4.8 shows an example from the feature selection process for the SELFBACK RCT I dataset. Right side of the figure presents the ten most correlated features used to build the CBR model for predicting one of the target outcomes *Numeric Pain Rating Scale (NPRS)* (input: baseline data, target: NPRS at follow-up 2). **np2** (eta-squared) is the squared correlation coefficient. Graph on the left shows the MAE (mean absolute error) variation with different sets of features in the corresponding CBR model for predicting *NPRS*.



	feature	np2
1	BT_pain_average	0.279547
2	Pain_worst	0.192200
3	Pain_1year	0.188000
4	EQ5D_pain	0.179000
5	BIPQ_symptoms	0.160972
6	EQ5D_anxiety	0.125000
7	BIPQ_pain_continuation	0.124556
8	BIPQ_life	0.116159
9	RMDQ	0.101490
10	EQ5D	0.091356

Figure 4.8: Correlation and Similarity-based Feature Selection. The x-axis presents the n-neighbours used for generating predictions and y-axis presenting the mean absolute error in the predictions for the entire dataset

4.4 Application of Machine Learning Methods on Healthcare Datasets

Based on the requirement imposed by the intended task, we looked into different methodologies and pipelines. In this section, we describe the methods applied in our research to develop learning models that can utilise the

intervention-based and population-based healthcare datasets for supporting clinical decision-making.

4.4.1 Case-Based Regression

Regression learning is a widely used application of CBR systems. Due to their transparency, case-based regression systems have gained popularity for complex healthcare datasets over conventional machine learning systems (Blanco et al., 2013). Rather than transforming the patient data into secondary representations as in tree-based learning or neural networks, CBR uses the stored knowledge directly to adapt and generate a solution for the new problem. Since a solution to a new problem is formulated by looking at similar past cases, this method has the advantage of transparency in the problem-solving process, thus making the system more comprehensible. The standard method for regression in case-based systems uses the solution S provided by n-nearest neighbours in the case base C determined using a similarity measure *sim* to provide a solution S_p for the new problem P. The solution S may be transformed using some transformation T in order to adapt and improve S_p and the number of neighbours n may vary. In this doctoral work, we applied the *similar problem, similar solution* principle of CBR and implemented Case-based regression models. Using the data-driven similarity modelling approach, we capture the inter-attribute and intra-attribute relationships within the dataset, determine predictive features using a tailored approach and adapt solutions from similar cases, thereby enabling the predictions to be easily explainable and transparent. Using a varied n-nearest neighbours approach and transforming the solutions of the similar cases by using a mean operator to adapt the solution to a new problem, we determine similar participant profiles in the HUNT4 dataset and predict patient-reported outcomes from the SELFBACK dataset.

- **Participant profiles: HUNT4**
  In this experiment to create activity profiles for the participants in the HUNT4 dataset, we first worked on developing a suitable similarity measure. We used the data-driven local similarity modelling approach, discussed in the last section, to model each physical activity attribute individually based on the attribute range in the dataset, followed by the global similarity where all the attributes were weighted equally. The main objective of the experiment was to determine similar

profiles and evaluate the quality of the similarity measures developed using the data-driven methodology. The quality of the retrievals was compared against that of the k-NN algorithm.

- **Predict Patient-reported outcomes: SELFBACK**
  Different CBR-regression models were implemented for predicting six hand-picked patient-reported outcomes at two follow-up time-points (at six weeks and three months) using baseline data in the SELFBACK RCT I dataset and evaluated their performance using an evaluation metric (MAE). The main objective of these experiments was multi-fold: (a) to build CBR systems for making individualized predictions of patient-reported outcomes at different follow-up time points using the baseline data in the SELFBACK datasets, and (b) to facilitate the determination of optimal feature subsets for predicting the patient-reported outcomes in RCT I dataset and finally, (c) evaluate the CBR models developed using the SELFBACK RCT II dataset as the external validation dataset.

4.4.2 Case-Based Clustering

Clustering is an unsupervised learning method used widely to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples (Han et al., 2011). When selecting clustering methods for healthcare datasets, it is important to consider that the aim is to partition the dataset into coherent clusters so that patients in any given cluster have a semantic relationship, more so than a syntactic one. A challenge with the most state-of-the-art clustering methods is the use of *knowledge-poor similarity metrics* or simple distance metrics such as hamming distance and Euclidean distance, among others. These metrics consider only the syntactic difference between two patient profiles, ignoring the coherence of each attribute, thus leading to insufficient estimation of the semantic similarity between them. It is necessary to build a similarity measure for these patient profiles that preserves their semantic relationship as much as possible and is also suitable for clustering the profiles since a clustering method will operate on the similarities between profiles in a given healthcare dataset. The similarity metric must allow the existing knowledge to influence the assessment of the similarity behaviour. Data-driven similarity offers a more versatile approach to handling clustering of

complex datasets by employing a knowledge-based approach (Müller and Bergmann, 2014). Focusing on the semantic similarity between attributes rather than the syntactic similarity, the collective influence of each variable's importance on the final (global) similarity score can improve the clustering quality by incorporating the existing knowledge in the dataset (Adam and Blockeel, 2015).

Knowledge-intensive Similarity-based Clustering

To ensure that each cluster is semantically coherent, we utilize the data-driven *similarity measures* discussed in the previous section 4.2 as the metric for clustering the dataset. The proposed approach, which takes inspiration from the working of the k-Means clustering algorithm (MacQueen, 1967; Lloyd, 1982), uses *knowledge-intensive similarity* as the metric for clustering. The algorithm extends the conventional approach of similarity in CBR by allowing the model to utilize the similarity measures aligned with domain expert knowledge and retains the semantic relationship between the cases. Algorithm 1 presents the *knowledge-intensive similarity-based* clustering algorithm for partitioning a case base into separate clusters.

Algorithm 1: Knowledge-intensive Similarity-based Clustering Algorithm

Input : case base C, number of clusters n
Output: n clusters
initialization: assign n random cases as centroids-$\{c_n\}$
Determine Cluster Membership
for *each case k in C* **do**
  compute $sim(k,c_j)$, $\forall j \in 1,...,n$
  assign k to most similar centroid
end
Update Cluster Centroids
for *each* c_j *in* $\{c_n\}$ **do**
  compute $meanSim_j = \frac{1}{|S_j|} \sum_{\forall k_i \in S_j} sim(k_i, c_j)$
  find case m in S_j such that
   $sim(m,c_j) \approx meanSim_j$
  assign m as the new centroid c_j
end
Repeat until centroids converge

S_j denotes the set of cases in cluster c_j.

Assigning n cases as centroids at random, the algorithm computes the

clusters using the similarity score of each individual case to each centroid, updating the centroids based on the average intra-cluster similarity (average similarity within a cluster). As our algorithm operates on the similarity score between each case and each centroid to determine its cluster membership, it is independent of the data type. As a result, one advantage of this clustering method is that it can easily be applied to different datasets other than just numerical, such as categorical or mixed datasets, which otherwise can be challenging when using the conventional clustering methods. Once the similarity measures are in place, one gets free from the trouble of taking care of the data types before applying the *knowledge-intensive similarity*-based clustering method. Such knowledge-based clustering systems would allow finding similar profiles in the same cluster and a similar yet significantly different profile in a neighbouring cluster to help increase the diversity of solutions.

Direct optimisation of algorithm 1 is an NP-hard problem and is, therefore, not always an option owing to high computational costs (Yang et al., 2016). Instead, a greedy approach can be used that surveys s-steps in the future to search for clusters with a higher average inter-cluster similarity (average similarity of all the clusters) and declares convergence if and only if no such clusters are found. To avoid falling into a local maxima, s needs to be large enough to accommodate the variation in the average inter-cluster similarity over multiple epochs. However, at the same time, s must be small enough to be computationally inexpensive for large datasets. This presents a challenge that is unique for each dataset.

As for estimating the optimal number of clusters n if it is not known before clustering a dataset, the most straightforward way is by plotting the sum of squared errors (SSE) against the number of clusters. With the increase in the number of clusters, SSE is expected to decrease, resulting in a *reversed elbow graph*, similar to that of an *elbow graph*for K-Means algorithm optimisation, but instead reversed. The reason is that the mean inter-cluster similarity is expected to increase with the number of clusters, thus resulting in a *reversed* elbow graph. The "elbow" of the graph would indicate the optimal number of clusters.

4.4.3 Conventional Regression

While CBR-based regression may be regarded as a lazy-learning approach since no computation takes place until a solution is required, conventional

supervised machine learning methods are quite the opposite. Conversely, the primary concern in supervised machine learning is the "*what*" rather than the "*how*", making conventional machine learning systems more concerned about accuracy than comprehensibility. That said, supervised machine learning methods have been widely used in developing prediction systems in the healthcare domain (Aldahiri et al., 2021).

- **Predict Patient-reported outcomes: SELFBACK**
  During the exploratory phase of the SELFBACK dataset, we investigated application of various machine learning regression algorithms to predict two patient-reported outcomes: *Pain Intensity* and *Workability Index*. The algorithms evaluated were: *Linear Regression* (Driver and Kroeber, 1932), *Passive Aggressive Regression* (Crammer et al., 2006), *Random Forest Regression* (Svetnik et al., 2003), *Stochastic Gradient Descent Regression* (Robbins and Monro, 1951), *AdaBoost Regression* (Freund and Schapire, 1997), *Support Vector Regression* (Boser et al., 1992), *XGBoost Regression* (Chen and Guestrin, 2016). Most of the algorithms gave similar performance, although two of them performed better than the rest: support vector machines and XGBoost. Subsequently, it was decided to only use these two algorithms for further experiments to build prediction models for six patient-reported outcomes selected from the RCT I dataset using features selected by the methodology described previously in section 4.3. We further evaluated the validity of these prediction models in a new (and final) experiment where we used the RCT II dataset as the external validation dataset and evaluated both the features selected as well as the prediction models. The main objective of these last two experiments was to evaluate the potential of the machine learning models and compare the performance of the support vector machines and XGBoost models to that of the corresponding CBR models (section 4.4.1).

Chapter 5

Research Results

This chapter presents an overview of the research papers, followed by a brief summary indicating how the papers relate to each other and finally, a summary of the research contributions. The research papers are included in full length in Part II of this thesis.

5.1 Overview of the Research Papers

The publications included in this thesis comprise three peer-reviewed conference papers, three journal papers and one peer-reviewed symposium paper. Six of these are already published, while the last one has been accepted for publication in a journal. The subsequent sections provide a *summary of main findings* and *core contribution* of each of the research papers including their *title*, *author name and contributions*, and the *publication venue*.

5.1.1 Paper A1

Title: Modelling Similarity for Comparing Physical Activity Profiles - A Data-Driven Approach

Author Names and Contributions: Deepika Verma, Kerstin Bach, Paul Jarle Mork

Verma was the main author and led the experiments, data analysis and wrote the paper. Bach provided expert knowledge for developing the methodology and contributed to the study design and experiments and writing the paper. Mork provided guidance on the results and analysis and feedback on

the paper.

Published In: International Conference on Case-Based Reasoning, pages 415-430, Cham, 2018. Springer. ISBN 978-3-030-01081-2

Summary of main findings: This paper describes the foundational work for this thesis, that is, similarity modelling for CBR systems. A data-driven methodology is presented for modelling the local similarities of continuous numerical attributes. The methodology utilises the information stored in the data, that is, *quartiles* of each attribute to adjust the behaviour of their polynomial similarity function such that the entire similarity range is encompassed. The utility of the proposed methodology is successfully evaluated by using the HUNT4 dataset to create a CBR model which can be used to compare physical activity profiles of the participants and comparing the performance of the thus created CBR model with an analogous k-NN model by retrieving the most similar profiles. Using MAE as the evaluation metric, it was found that the profiles retrieved by the CBR model had the least error as compared to their k-NN counterpart. This effect is due to the CBR model's ability to capture the variation in each attribute owing to the data-driven local similarity modelling, which is not the case for the k-NN model. Furthermore, while the contents of this paper are domain-specific, the methodology presented is general and can be applied to other datasets with numerical attributes.

Core contribution: The paper contributes a data-driven methodology for modelling local similarity measures for continuous numerical attributes to build CBR systems. The methodology is not domain specific and ensures that the intra-attribute variance for each of the attributes in a given dataset is preserved, and the entire similarity range [0,1] is utilised when modelling the local similarity measures.

5.1.2 Paper A2

Title: Similarity Measure Development for Case-Based Reasoning – A Data-Driven Approach.

Author Names and Contributions: Deepika Verma, Kerstin Bach, Paul Jarle Mork

Verma was the main author and led the experiments and wrote the paper. Bach provided supervision for the study design and experiments and provided feedback on the paper. Mork provided feedback on the paper.

Published In: Norwegian Artificial Intelligence Society, pages 143–148, Cham, 2019. Springer. ISBN 978-3-030-35664-4

Summary of main findings: This short paper builds on the data-driven similarity modelling approach presented in Paper A1 and empirically demonstrates the generalizability of methodology by using an open-source dataset with an application domain other than healthcare as an example. The data-driven similarity modelling approach was as effective for the new dataset for estimating similarity among the users as for the participants in the HUNT4 dataset in paper A1.

Core Contribution: The paper demonstrates the applicability of the data-driven approach for modelling local similarities, presented in paper A1, on datasets with application domains other than healthcare.

5.1.3 Paper B

Title: Clustering of Physical Behaviour Profiles using Knowledge-intensive Similarity Measures

Author Names and Contributions: Deepika Verma, Kerstin Bach, Paul Jarle Mork

Verma led the research and paper writing and was the main author. Bach contributed towards the study design, evaluation of experiments and writing the paper. Mork provided general supervision for the research and feedback on the paper.

Published In: International Conference on Agents and Artificial Intelligence, Volume 2, pages 660–667. INSTICC, SciTePress, 2020. ISBN 978-989-758-395-7

Summary of main findings: This paper presents a clustering algorithm for CBR models that uses the similarity measure as the metric for clustering the case base into coherent clusters such that intra-cluster similarity is maximized. The similarity measure is developed using the data-driven methodology presented in Paper A, and is the reason the algorithm is called *knowledge intensive similarity*-based clustering algorithm since the similarity measure used for clustering the case base takes into account both the knowledge contained in the dataset and the expert knowledge required to develop the similarity measures. The algorithm uses the knowledge-intensive similarity to partition the case base into separate clusters while preserving the intra-cluster semantic similarity. The algorithm gets evaluated using the HUNT4 dataset by analysing the generated clusters. The evaluation

addresses the maximum optimal number of clusters in the given case base and coherency of the clusters obtained using intra-cluster mean similarity as the evaluation metric. On comparing the results with the k-Means clustering algorithm, the proposed clustering algorithm generated more coherent clusters with higher intra-cluster similarity amongst the cases. While the knowledge used for creating the CBR model and evaluation of the proposed clustering algorithm in this paper is domain-specific, the clustering method itself is general and can be used on datasets of other domains.

Core contribution: The main contribution of this paper is a knowledge-intensive similarity-based clustering method that uses data-driven similarity as the metric for clustering a case base to maximise the intra-cluster similarity and increase the diversity of solution.

5.1.4 Paper C

Title: Exploratory Application of Machine Learning Methods on Patient Reported Data in the Development of Supervised Models for Predicting Outcomes

Author Names and Contributions: Deepika Verma, Duncan Jansen, Kerstin Bach, Mannes Poel, Paul Jarle Mork, Wendy Oude Nijeweme d'Hollosy

Verma led and coordinated the paper writing and was the main author. Verma and Jansen conducted the experiments and contributed to writing the paper. Bach and d'Hollosy contributed to the study design, paper writing and provided supervision for experiments. Mork and Poel provided general supervision for the research and feedback on the paper.

Published In: BMC Medical Informatics and Decision Making, 22(227), 2022. ISSN 1472-6947.

Summary of main findings: This paper was a joint effort with researchers in the Netherlands, wherein exploratory experiments were conducted to explore traditional machine learning methods on two datasets consisting of PROMs collected from patients with neck and/or low back pain in separate clinical trials. The main aim of this paper was to investigate whether and to what extent different machine learning methods can predict and classify patient-specific "target" outcomes based on patient-reported data. Feature selection strategies were discussed to determine which patient-reported measurements are most descriptive of the target outcomes. The prediction models built for the first dataset wherein the baseline measurements recorded for the associated target outcomes gave promising

results. Moreover, the baseline measurements were the most useful predictor of the associated target outcome. On the other hand, the classifiers built for the second dataset performed poorly due to the lack of appropriate/valuable predictors of the target outcome. The results presented a promising potential for using machine learning for predicting or classifying PROMs, provided that PROMs hold enough predictive power.

Core Contribution: Emphasis on the utility of machine learning methods for the exploration of PROMs and determination of supervised machine learning methods suited for the task of predicting outcomes from PROMs datasets.

5.1.5 Paper D

Title: Using Automated Feature Selection for Building Case-Based Reasoning Systems: An Example from Patient-Reported Outcome Measurements

Author Names and Contributions: Deepika Verma, Kerstin Bach, Paul Jarle Mork

Verma led the research and paper writing and was the main author. Bach contributed towards the study design, evaluation of experiments and writing the paper. Mork provided general supervision for the research and feedback on the paper.

Published In: British Computer Society, Specialist Group on Artificial Intelligence, pages 282–295, Cham, 2021. Springer. ISBN 978-3-030-91100-3.

Summary of main findings: This paper presents a two-fold hybrid approach for selecting features from a subjective intervention-based healthcare dataset to build CBR model(s) that can predict patient-reported outcomes. In the two-fold approach, correlation is used first to filter out the most correlated features, followed by building a CBR model using the data-driven knowledge modelling presented in Paper A to derive the optimal feature subset. The optimal feature subset is decided based on the trade-off between the number of features used and the error (MAE) in predicting the target outcome. The quality of features selected using this two-fold approach is compared with another method, permutation feature importance using XGBoost as the base regressor, for deriving feature selection. Based on the prediction error (MAE), it was found that the results produced by the CBR prediction models built using features selected by either of the methodologies are similar. However, considering the time and effort trade-off, the latter method

may be the more preferred option, notwithstanding the utility of the former. From a clinical perspective, the baseline measurements of the associated target outcomes were their most important predictors, a result supported by similar findings in other studies in the literature. Further, the performance of the CBR prediction models was comparable to the traditional machine learning regression models—Support Vector Machines and XGBoost.

Core Contribution: An approach for selecting optimal feature sets for various outcomes in a PROMs dataset for the purpose of developing CBR models to predict the outcomes as well as a comparison of the predictive performance of CBR models and two supervised machine learning algorithms.

5.1.6 Paper E

Title: Application of Machine Learning on Patient-Reported Outcome Measurements for Predicting Outcomes: A Literature Review

Author Names and Contributions: Deepika Verma, Kerstin Bach, Paul Jarle Mork

Verma led the literature review and paper writing and was the main author. Bach provided guidance for the review design and contributed to writing the paper. Mork provided expert domain knowledge for the study design and supervision for the research and provided feedback on the paper.

Published In: MDPI Informatics, 8(3), 2021. ISSN 2227-9709.

Summary of main findings: This paper is a literature review that summarises the recent trends in the application of machine learning methods on PROM datasets for predicting clinical outcomes. The review analyses the articles based on their year of publication, domain of intervention, length of prediction (in terms of weeks or months, or years), source of the dataset used, use of external validation dataset, feature selection strategy, and the machine learning methods applied. All the included articles were published in peer-reviewed journals and could be broadly categorized into five intervention domains which include post-surgical improvements or limitations, depression, pain management, hospital readmission, and oral health. Post-surgical and depression interventions were the dominant theme in the included articles. An emerging trend was discovered with an increase in the use of machine learning methods for patient-reported datasets for feature selection and patient-specific outcome prediction. Ensemble and linear methods were the most commonly used methods both for selecting features and predicting outcomes. One of the major gaps identified was the lack of ex-

ternal validation, as only four of the fifteen included articles included an external validation of the models. This gap could be due to either lack of or unavailability of an appropriate external validation dataset, as complete and readily available datasets for research purposes are rare to find and is one big challenge in themselves. Another gap was the inconsistency in reporting essential elements of the development of machine learning models such as the selected features, method of hyperparameter tuning, and hyperparameters used, which makes result reproducibility and further research a challenge.

Core Contribution: A review of the existing state-of-the-art for application of machine learning methods on PROMs datasets to predict individualised patient-reported outcomes.

5.1.7 Paper F

Title: External Validation of Prediction Models for Patient-Reported Outcome Measurements collected using the SELFBACK Mobile App

Author Names and Contributions: Deepika Verma, Kerstin Bach, Paul Jarle Mork

Verma led the experiments and paper writing and was the main author. Bach provided guidance for the study design and approved the experimental analysis and contributed to writing the paper. Mork provided general supervision for the research and provided feedback on the paper.

Published In: Accepted for publication in *Elsevier International Journal of Medical Informatics*.

Summary of main findings: This paper presents an external validation of the prediction models built during an experimental work to select relevant features for and predicting individualised patient-reported outcomes using baseline PROMs, the results of which were presented in Paper D. The prediction models were trained using a dataset that consisted of PROMs collected in the RCT I for the evaluation of the SELFBACK mobile application. Since internal validation on smaller datasets is generally not optimal enough for evaluating generalisability of the models, external validation is essential. The dataset used for external validation consisted of the same type of PROMs as in the training dataset, collected in the RCT II as a part of evaluating the efficacy of the SELFBACK mobile application. Overall, the predictive power was low, except for prediction of one of the outcomes. The results indicate that the models show ability to generalise and predict out-

comes for a new dataset and highlight the need for external validation in healthcare-oriented studies for the further development of patient-centred healthcare systems.

Core Contribution: The paper contributes toward external validation of individualised outcome prediction models built for PROM datasets to facilitate further development of CDSS.

— — —

With paper A1, we established a data-driven methodology that can be employed to model attribute similarities for building CBR systems and demonstrated our results on the HUNT4 dataset. Paper A2 later builds on paper A1 to demonstrate the generalisability of the data-driven methodology on a dataset of an unrelated domain. The encouraging results from these two publications subsequently led to paper B where we used the data-driven similarities as the clustering metric, documenting a favourable effect on the HUNT4 dataset. With paper C, we moved onto the SELFBACK datasets and did exploratory work with conventional machine learning methods for PROM datasets, which acted as the basis for subsequent research with the SELFBACK dataset. Building on the findings of the previous papers (A1-C), papers D and F build and evaluate CBR prediction models and feature selection strategies that include data-driven similarities of CBR models and permutation feature importance of ensemble methods. The papers also present a performance comparison of the case-based prediction models with conventional machine learning models. Paper E summarised the existing state-of-the-art forming the basis for papers C, D, and F. The following section describes how the papers relate to the research questions and the main contributions from the research question.

5.2 Summary of research contributions

This section summarizes the main research contributions of this thesis. The summary below references the related publications pertaining to each of the three research questions, as set out in the introduction chapter (section 1.2) individually.

5.2.1 Research Question 1: How to measure similarity among different individuals based on their objective and subjective measurements?

- Section 4.2 describes the data-driven methodology for developing the similarity measures for building CBR models using healthcare datasets. By utilising the spread of the values in each attribute and defining strong initial values with the help of strategic markers (quartiles for numerical attributes), similarities measures can be developed which can be used to measure similarity among individuals based on their objective or subjective measurements.

- Paper A1 and B address this research question. Paper A1 presents the data-driven methodology for modelling local similarities of attributes in the HUNT4 dataset containing objective physical behaviour measurements from participants of a cohort study. By using the methodology on the objective measurements, we built individual physical activity profiles in a CBR system and using the constructed similarity measure, measured the similarity among the individual profiles in the case base. In paper B, we used the methodology presented in paper A1 to build the individual physical activity profiles on an extended HUNT4 dataset, measured similarity among the individual profiles and further utilised the constructed similarity measure to cluster the profiles into coherent clusters. We demonstrated that the data-driven similarity functions we build can be used to measure similarity among different individuals.

- Paper A2 relates to this research question. In this paper, the data-driven local similarity modelling methodology presented in paper A1 is applied on an open-source dataset of a different application domain, to estimate similarity between individual profiles. By doing so, we also demonstrated the generalizability of the methodology.

- Paper D and F address this research question. In both these papers, individual patient profiles of the participants in the SELFBACK RCTs are built based on their subjective measurements using the data-driven similarity modelling methodology described in paper A1 (the paper describes the methodology only for the numerical attributes) and in section 4.2.

- The proposed case-based clustering approach in paper B is the first in the literature utilising purely data-driven similarities for clustering activity profiles of participants of a population-based cohort study. Smyth (2019) previously proposed a case-based approach to recommend pacing plans for marathon runners, albeit without any clustering.

5.2.2 Research Question 2: How can machine learning methods be applied to subjective patient-reported datasets to facilitate individualized patient-reported outcome predictions?

- Section 4.3 describes the feature selection strategy for selecting optimal feature set from a PROM dataset that can best predict a selected patient-reported outcome. The two-fold hybrid feature selection approach described later in the section helps determine a set of features targeted specifically for building CBR prediction models, though can be used for general machine learning as well. While the other approach, permutation feature selection, provides feature sets that are more general.

- Papers C and D address this research question. Paper C explores a number of machine learning algorithms with the goal of determining the most suited algorithms for the task of predicting individualized patient-reported outcomes based on patient-reported measurements in PROMs dataset. The paper presents the results of feature selection strategies and subsequent prediction models for predicting pain-related outcomes, thereby contributing towards feature selection and model selection for predicting pain-related outcomes. Paper D builds on paper C and presents the two-fold feature selection approach and its utility in comparison to a traditional feature selection approach.

- Paper F relates to this research question. The paper broadly addresses the challenge of external validation of models for predicting patient-reported outcomes to verify the generalisability of such models. Using an external dataset, we evaluated the generalisability of the prediction models presented in paper D and by doing so, also emphasised the

importance of external validation of prediction models intended for facilitating clinical decision support.

- The CBR models developed and presented in our work are the first where CBR has been applied exclusively on PROM datasets and for predicting patient-reported outcomes using PROMs alone. We also demonstrated that data-driven similarities can be utilised to successfully drive feature selection with CBR models.

5.2.3 Research Question 3: What are the state-of-the-art of machine learning methods for investigating patient-reported outcome measurement datasets?

- This research question is addressed by paper E which provides an overview of the existing literature on application of machine learning methods on PROM datasets during the last decade. The paper summarizes findings from fifteen articles and presents a comprehensive analysis of several aspects such as the intervention domain, time period between measuring the predictor(s) and the outcome, feature selection methods, machine learning methods, optimization techniques, and external validation, among others. Ensemble and linear methods were the most commonly used machine learning methods on PROM datasets for feature selection and outcome prediction.

- No published articles were found that concern the use of CBR systems on PROM datasets, thereby highlighting a gap in the existing literature. This gap presents a unique opportunity to delve into the utility of CBR systems for subjective patient-reported clinical datasets, and our work provides a basis for further exploration of this area of CBR application.

- Ours is the first literature review focusing solely on the state-of-the-art concerning applied machine learning on PROMs datasets for predicting individualized patient-reported outcomes.

Chapter 6

Discussion

The research for this thesis was motivated by the overarching goal of advancing research in the utilisation of case-based methods for healthcare datasets to underpin and facilitate further development of patient-centred care through CDSS. In our investigation into the research questions, we implemented supervised and unsupervised case-based learning methods for both population-based and intervention-based datasets that can facilitate decision-making in tailoring physical activity plans for population subgroups and monitoring pain-related outcomes of patients with LBP or NP. Our research started with analysing the objective physical behaviour measurements of participants in the HUNT4 dataset to get insights into the activity distribution and characteristics of the population sample in the dataset. We created activity profiles of the participants, modelled the similarity among them using a data-driven approach, and built a CBR model in *myCBR workbench* to determine similar activity profiles (published in paper A1). Based on the results so far, we decided to utilise the data-driven global similarity in the CBR model to cluster a case base with more participant activity profiles from the HUNT4 study to identify groups with similar physical behaviour that can facilitate tailoring of physical activity plans (published in paper B). We then moved on to explore the subjective measurements collected in the SELFBACK project and analysed the utility of different machine learning methods in investigating the predictive potential of PROMs. Doing so can support several aspects of clinical decision-making process such as identifying baseline PROMs with predictive power to condense clinical questionnaires, moderating the information burden on clinicians, and monitoring the follow-up of patients to identify those at a

higher risk of a worse outcome. Based on the existing literature in the application domain (later published in paper E), we examined several conventional machine learning methods to identify predictors of two patient-reported outcomes (in paper C). Based on our findings, we experimented with a hybrid approach for feature selection which included utilising a combination of data-driven similarities (modelled for the patient profiles in the CBR model in myCBR workbench) and correlation to determine predictors of patient-reported outcomes and built case-based prediction models (published in paper D). Finally, we evaluated the validity of both the predictors and the prediction models using an external validation dataset (paper F).

While the research done in our work has been driven towards facilitating the development of CBR-based decision support systems, the methods applied and developed in this thesis may not be limited in their application to any particular domain. We hypothesise, based on the nature of the approaches undertaken and methods developed, that our work in this thesis can be generalised and be of use in other domains. That said, the work presented in this thesis has several limitations, and more work is necessary to address these limitations and carry forward the research. In the sections ahead, we discuss the limitations of this thesis, further research possibilities that may be addressed in future work, and finally, conclude this thesis.

6.1 Limitations

Similarity The global similarities developed for the CBR models have some limitations regarding completeness and weight distributions. We mainly covered the numeric and symbolic similarity measures in our research. More complex data types can likely not be modelled by the proposed data-driven method. Among the CBR models developed in our work, all the features were weighted equally. The features may be weighted differently since every feature may not contribute equally to a model, which might have led to a different performance. However, we did investigate several different methods to determine feature weights and estimated the model performance, though it was found to not be any better than with equally weighted features. Nevertheless, there may be other approaches for feature weighing that may yield better results.

Adaptation The case-based prediction models for the SELFBACK datasets have a limitation that a rudimentary transformation approach has been used for adapting the solutions—an average over the n-nearest neighbours—for the mere reason that for any given query, there is no guarantee to find a case that fits the problem description. Advanced adaptations would require the involvement of domain experts or revision steps to check for the validity of the adaptations. Another limitation concerns the number of cases n chosen for reuse, which was not fixed and varied significantly due to high variability in the reported outcome among the similar profiles. Other advanced adaptation approaches may be explored in future studies to address these limitations.

Algorithm convergence and Optimisation Firstly, the knowledge intensive clustering algorithm described in section 4.4.2 may not yet produce an optimal convergence. Since optimisation of such an algorithm is an NP-hard problem, a greedy approach was employed that considers *s-steps* in the future before declaring convergence. The step size s may vary depending on the dataset and application requirements and is, therefore, not a fixed entity. There might, however, be other better or more optimal ways to achieve convergence. Secondly, grid search was employed for optimising the hyperparameters of the machine learning algorithms used in the papers. Grid search may not be the most optimal method of tuning the hyperparameters, both in terms of time and hyperparameter. Other approaches such as Bayesian optimisation may produce better results at a fraction of the time cost.

Physical Activity The possibility of making tailored physical activity plans cannot yet be analysed. There is a need for a holistic 24-h analysis of the physical behaviour data and associations with various health outcomes to capture more relevant information that may be necessary to get a better understanding of their dynamics.

Validity The validity of the features selected for the corresponding target outcomes is difficult to assess since most clinicians have a hard-time hand-picking meaningful features from the entire pool. While the features chosen are statistically relevant, whether or not they are informative in a clinical setting could not be assessed. Furthermore, features deemed important in

one model may not necessarily be equally important in another model. As such, variation in the choice of features selection strategy and the features eventually used in the models will inevitably produce different results. To this end, further research is necessary. Regarding external validity, which concerns the generalization of the prediction models, our experiments in paper F presented promising results, indicating some degree of generalisability. However, the results were not entirely exemplary, and there is likely some scope for improvement in further research.

Literature Search The literature review may have some limitations concerning the choice of search string and database used to identify relevant articles. Some important works may have been missed due to differences in keywords used in the published articles and databases. Furthermore, the literature search did not include CBR altogether and, therefore, may not represent a comprehensive overview of the extent of literature available in the application domain.

6.2 Future Directions and Conclusion

The following future directions could provide an avenue for further investigation as an extension to the doctoral research presented in this thesis.

Exploring other methods Further research could involve investigating other schemes for similarity measure development, case representation, case adaptation and clustering in CBR systems. For the clustering algorithm (presented in section 4.4.2), other criteria or methods for convergence must be explored and analysed. Alternative methods for feature selection and prediction should also be explored and compared with the results in this thesis to get better insights into the potential of the SELFBACK datasets. It may also be of interest to compare our results with new prediction models built using features selected by the domain experts to get a better understanding of how they compare with the ones chosen by the machine learning methods. Another possibility to explore would be to cluster the SELFBACK datasets using the clustering algorithm (section 4.4.2) and then apply feature selection cluster-wise, followed by patient-specific outcome prediction, similar to the approach taken by Chekroud et al. (2017). Further, other approaches for

case adaption may be explored and analysed to determine their suitability for the task.

Evaluation studies for validation The models developed in this doctoral research work are still in the early stages, and additional work is necessary for thorough evaluation. Further studies may involve designing experiments with larger sample sizes and perhaps using other datasets collected in different settings to address the validity of the results.

Involving the clinicians and patients To evaluate the usefulness and benefit of this research work, future work may involve clinicians and patients. Collaborative research with domain experts would be necessary to investigate the adequacy of the chosen features for clinical decision-making. And, the clinicians would be able to assess the potential of the developed methods and systems for decision support in the clinical settings. Engaging the patients in the decision-making process might help in improving their acceptance of the system and possibly, adherence to the treatment (Bitton et al., 2014). It will also allow to validate and improve on the instrumental concepts, analyze the added value of the tool for the primary users, and elucidate more detailed policies for designing and implementing CDSS.

Prototyping and testing The proposed clustering approach in Paper B is motivated by multiple elements of existing research for identification of groups with similar patterns of physical activity (Marschollek, 2013) and using that knowledge along with a case-based approach to recommending activity plans (Smyth, 2019). That said, our work in this thesis is limited to the steps leading up to building a case-based approach for creating physical activity plans based on individual profiles, and thus, requires more work that may be addressed in future studies to implement and test the concepts, ideally in collaboration with the domain experts. Future research could also include refining and extending the prediction models for SELFBACK after thorough validation to implement a working prototype that can be tested in real-world settings in collaboration with clinicians and patients. This step would be essential to assess the utility and impact of such decision support systems on the care of patients with low back pain and bring us one step closer to incorporating them in clinical practice.

—

In conclusion, the work in this doctoral research began with the basic building block of devising a methodology for the development of similarity measures to build CBR models and ventured gradually towards implementing case-based learning models for identifying groups with similar physical activity patterns in population-based studies and forecasting individualised patient-reported outcomes based on PROMs in intervention-based studies using the same building blocks. The lack of any literature applying CBR for only PROMs also highlights a gap that may be addressed in future studies. Further, our results have demonstrated the expedience of the data-driven similarity measures developed, both for case retrieval and clustering, and the potential of using a case-based approach for subjective healthcare datasets, providing a proof of concept suggesting that CBR systems accord an ideal platform for harnessing the knowledge stowed away in healthcare datasets. The results also indicate that a close liaison between patients (through the data), clinicians, and case-based methods can bring about a better understanding of patient-centred care and, thus, provides grounds for further research and development of more transparent, evidence-based decision support systems for healthcare settings.

Part II

PUBLICATIONS

Modelling Similarity for Comparing Physical Activity Profiles - A Data-Driven Approach

Deepika Verma, Kerstin Bach, Paul Jarle Mork

Abstract

Objective measurements of physical behaviour are an interesting research field from the public health and computer science perspective. While for public health research, measurements with a high quality and feasible setup is important, the analysis of and reasoning about the data is what we will present in this work. Our focus in this work is the comprehensive representation of physical behaviour throughout consecutive days and allowing to find sub-groups in the population with similar physical activity levels.

We have a unique data set of 4628 participants wearing tri-axial accelerometers for six days and will present a case-based reasoning (CBR) system that can find and compare similar activity profiles. In this work, we focus on creating a CBR model using myCBR and do initial experiments with the resulting system. We will introduce a data-driven approach for modelling local similarity measures. Eventually, in the experiments we will show that for the given data set, the CBR system outperforms a k-Nearest Neighbor regressor in finding most similar participants.

A Introduction

Physical inactivity and poor sleep are considered global health problems (Kohl et al., 2012; Raitakan et al., 1994) that contribute substantially to poor health and premature mortality. It is estimated that physical inactivity is responsible for about 9% premature mortality (Lee et al., 2012), which is similar to the effect of smoking (Wen and Wu, 2012) and obesity (A et al., 2017).

CBR has become more popular over the last few years, especially in an area where continuous measurements become more and more available (Canensi et al., 2017; Plis et al., 2014). It offers a way for abstracting and transferring specific domain expert knowledge into a self-explanatory and user-friendly tool, which can be used to generate solutions for problems ranging from simple daily life tasks to complex issues (which otherwise necessitate expert help), with an appropriate reasoning behind them. Not only is it being applied for finding similar cases to provide solutions, but also for the classification of medical (Yao and Li, 2010; Campillo-Gimenez et al., 2013) and activity data (Uddin and Loutfi, 2013). In Uddin and Loutfi (2013),

the authors propose a CBR method to classify different physical activities of elderly based on their pulse rate.

In this paper, we focus on the knowledge engineering process of creating a CBR model and present a data-driven approach for modelling local similarity measures for physical activity data in the myCBR workbench (Bach and Althoff, 2012; Stahl and Roth-Berghofer, 2008). We will show in our experiments that a CBR system comparing physical activity profiles is less erroneous than a k-Nearest Neighbour (k-NN) regressor model. In our experiments, both approaches are used to find groups of similar activity profiles and their performance is evaluated statistically. The second contribution of this paper is a method for modelling the local similarity measures utilizing data driven methods. We will showcase how a data set can lead to strong initial definitions for numerical value ranges and therewith easen and stratify the knowledge modelling process.

The remaining of this paper is divided into sections as follows: in section B, we discuss related work on reasoning about physical activity behaviour using various approaches within machine learning and artificial intelligence. In section C, we discuss the importance of objective measurements of physical activity behaviour from both public health and computer science perspective. Section D is dedicated to similarity modelling for the data set in myCBR. In section E, we present the experiments performed to evaluate the CBR model generated and compare it with that of k-NN model. Section F and G are for discussion and conclusion respectively.

B Related Work

The amalgamation of sensors, Internet of Things (IoT) and Artificial Intelligence (AI) provides a unique opportunity not only for health researchers, but also for AI researchers to perform objective measurements and utilize raw data recordings to generate physical activity profiles of a large number of participants and determine similar physical activity profile groups. With the help of AI techniques, it is possible to perform objective analysis of sensor data stream to not only identify different physical activities uniquely (Bulling et al., 2014; Arif and Kattan, 2015; Willetts et al., 2018), but also find out groups of similar activity profiles. Finding and clustering similar physical activity profiles is crucial in facilitating the understanding of health and activity characteristics of a population and identifying different activity

phenotypes[1]. In Marschollek (2013), the author proposed an ATLAS index to cluster and identify four activity phenotypes using NHANES[2] data set. Similarly, in Willetts et al. (2018), authors proposed a statistical machine learning model to identify different sleep and physical activity phenotypes. Further, the authors in Howie et al. (2018) apply latent class analysis to identify five different activity phenotypes among young adults in a cohort study where data was collected using hip-worn accelerometers for seven days. Our long term goals and target data are similar to these studies, however the approach differs slightly.

Similar to the preference-based CBR framework presented by Hüllermeier and Schlegel (2011), we are presenting a framework for modelling local similarity measures based on the data set available. Therewith we can tailor each similarity measure to the application domain. In the continuation of their work Abdel-Aziz, Strickert and Hüllermeier (Abdel-Aziz et al., 2014) show that the data distributions and distances in data sets can be used for learning similarity measures. While the authors focus on learning preferences, we show with the work presented here that the data-driven view can be carried over to general knowledge engineering tasks. Using a data-driven approach for automatic similarity learning and feature weighting has been presented by Gabel and Godehardt (Gabel and Godehardt, 2015). In their work they trained a neural network to induce local and global similarity measures. While we are not automatically assigning the similarity measures, we also use existing cases to derive them. In Smyth and Cunningham (2017), the authors explored a case-based approach for recommending 5km times for marathon runners in order to achieve their personal best. The approach they apply is similar to the one presented in this paper as they use timing profiles as basis for the similarity-based assessment. In a slightly different approach, Sani et al. (2017) explored using k-NN for detecting physical activities from wrist worn sensors. In their work they show that applying k-NN for detecting movement patterns is very successful for creating personalized models. Even though the approaches differ, our work is similar in terms of comparing physical activity profiles with raw data coming for accelerometers.

[1] *https://www.biology-online.org/dictionary/Phenotype*
[2] *https://wwwn.cdc.gov/nchs/nhanes/default.aspx*

C Physical Activity Analysis for Public Health Application Scenarios

Regular physical activity is important for people of all age groups, including the elderly. It can significantly reduce the risk of various health problems such as stroke, diabetes, various types of cancer, depression, as well as hypertension and improve bone and muscle health[3]. Physical inactivity is one of the most important public health problems of this century and has a strong negative impact on the physical and mental well being of an individual. It is estimated that about 23% adults and 81% adolescents globally are physically inactive. The figures are alarmingly high for adolescents. Moreover, being physically active is not just about moving around in the house or walking at a slow pace, they must include some form of Moderate to Vigorous Physical Activity (MVPA) such as brisk walking, dancing, running, cycling, or moving/lifting heavy load.

Over the last few years, researchers in public health domain have moved rapidly from using self-reported subjective activity data to objectively measured activity data with the use of body-worn sensors (Arif and Kattan, 2015; Lee and Shiroma, 2013; Li et al., 2017). Not only are the sensors a more viable option due to the simplicity of extracting and utilizing raw data, but also eliminate the problem of bias due to self reporting (Prince et al., 2008; Lagersted-Olsen et al., 2013), which has been a major concern among researchers as it leads to inaccuracy and uncertainty. Moreover, the accelerometers directly measure the subject's physiology motion status to indicate the motion pattern within a given time period, which is helpful in activity recognition and are much more energy efficient.

The physical activity data used for this work is primarily based on accelerometer data collected during the HUNT4[4]cohort study. The Nørd-Trøndelag Health Study (HUNT)[5] in Norway is one of the largest health studies of its kind. The study consists of a large amount of health data collected through questionnaires and clinical examinations during three intensive previous studies (HUNT1 1984-86, HUNT2 1995-97 and HUNT3 2006-08). In the ongoing study HUNT4 (2017-19), each participant is offered to participate in the objective measurements data collection. If accepted,

[3] $http://who.int/features/factfiles/physical_activity/en/$

[4] $https://www.ntnu.no/hunt4/$

[5] $https://www.ntnu.no/hunt/$

they are fitted with two wearable tri-axial accelerometers, placed at their thigh and lower back, which are used to collect activity data for one week. The raw sensor data is then classified into 17 different physical activities using Support Vector Machines (for the synchronization of sensor data) and Random forest classifiers (for the prediction of activity classes). Afterwards, these activities are grouped into six main physical activities: lying, sitting, standing, walking, running, cycling, which is the basis data set for our work[6].

By determining the variation among participants in different activity clusters through similarity, it is possible to provide activity recommendations to less active profiles in order to make them more active. Every person has different activity characteristics and finding a group of activity profiles most similar to that person with respect to the duration of every activity is a challenging task and we aim to address this task using Case-Based Reasoning (CBR), because it offers the flexibility and transparency in its reasoning process.

D Data-driven Knowledge Modelling

In this section, we explain how we implement a CBR system that can be applied to find and compare similar activity profiles from objectively measured population data. We are using the local-global-principle (Richter, 1995) for creating similarity measures and thereby build a knowledge model that tailors the similarity measure for each attribute. Once the local similarity measures are defined, we continue to use weighted sum for defining the global similarity.

While the HUNT4 data set is unique in the world, the challenges for utilizing it for developing a CBR system are very common such as the identification of suitable data set context for the problem at hand, definition of initial similarity measures, representation of cases and determination of valuable cases for populating the casebase. In this work we will introduce a method for utilizing a given data set to model similarity measures. Further we will take into account the effect of growing case bases and show a methodology that can help to visualize and understand how a CBR system learns.

This section is further divided into subsections as follows: First, we describe how we populate the casebase and generate cases in the developed case representation. Second, we describe our data-driven approach to model

[6] *Since the study is ongoing, we have used the data available by March, 12 2018.*

the local similarity measures for the numerical activity attributes. Once the model is in place, we then query the casebase and compare the most similar activity profiles retrieved.

D.I Case Generation

Developing a case representation is the first part of the system development. Depending on the domain and the available data this can be a challenging process on its own (H. El-Sappagh and Elmogy, 2004; Bergmann et al., 2005; Khamparia and Pandey, 2017). For our application domain we utilize the pre-processed HUNT4 data. While HUNT4 collects a very comprehensive set of data, we are only focusing on the objective measurements. The sensor data is collected over a period of seven days per participant and the overall data collection in the cohort stretches over 18 months, starting from the autumn of 2017 until February 2019. It is an ongoing study and until March 2018, data for 17409 participants has been automatically classified and for each participant aggregated into the six main physical activities. In Table 1 we present the description of the six activity types used in our data set.

Activity	Description
Lying	The person lies down
Sitting	When the person's buttocks is on the seat of the chair, bed or floor
Standing	Upright, feet supporting the person's body weigh
Walking	Locomotion towards a destination with one stride or more
Running	Locomotion towards a destination, with at least two steps where both feet leave the ground during each stride
Cycling	The person is riding bicycle

Table 1: Activity Descriptions

Each participant is fitted with two tri-axial accelerometers, AX3 Axivity[7], one on the thigh and second on lower back. The sensors are used to detect vibrations, movement and orientation changes in the three axes. The

[7] *https://axivity.com/downloads/ax3*

sampling frequency of the sensors is set at 50Hz. After the participant has worn the sensors for seven days, they are returned to the HUNT research center where the raw data is downloaded, extracted and classified using Support Vector Machines and Random Forest algorithms. The resulting data set contains the H4ID (unique ID for each HUNT4 participant), number of minutes of each different activity, the date and day of the week in a csv file.

When preparing the data for the CBR system, we further process it by removing the records where we assume the sensor was taken off or the prediction failed. Those are very long times of the same activities. Records are removed based on the following criteria:

- sum of all the activities for a single record exceeds 1440, which is the total minutes in a day
- records containing zero minutes for lying, sitting, standing and walking
- data set for one participant has less than seven days of data

Eventually, we chose to keep records where exactly six days of data per H4ID was present, while the rest of the records were removed. For each unique H4ID, the total minutes of each activity were summed up for six days. We experimented with different knowledge representations including mean, maximum and sum of duration of each activity per H4ID and found the sum representation to suit best since it captures the overall physical behaviour of the participants over the days as well as the variance of the similarity measure over its' entire range. At this point, after pre-processing, the data set contains 4628 rows, each record containing sum of each activity over six days for a single participant. Table 2 gives a brief account of the data set.

	Lying	Sitting	Standing	Walking	Running	Cycling
count	4628	4628	4628	4628	4628	4628
mean	3090.49	3322.82	1401.22	790.67	6.86	26.45
min	7.35	253.25	56.50	1.55	0	0
max	7513.80	7846.10	4247.10	2101.65	172.70	719.10

Table 2: Data set Statistics

Cases are populated from the previously described data set by loading into the previously defined case representation using the myCBR tool. A

single case in myCBR is represented as shown in Fig. 1, where *Participant* is the name of the concept which consists of six attributes namely *cycling, lying, running, sitting, standing and walking.*

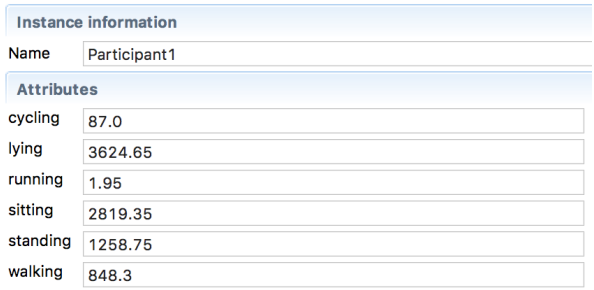


Figure 1: Case representation in myCBR

D.II Data-driven Similarity Measures Development

The local-global-principle requires that both types of similarity measures, the local one on the attribute level and the global one on the conceptual need to be defined.

Modelling the local similarity measures for different attributes in myCBR can be challenging as researchers have to balance the input from the domain experts and the available data. Having criteria which can lead the knowledge modelling process is helpful for both parties. We therefore suggest to make use of the existing data in this process. As we assume that the collected data set covers the scope of what type of problems (cases) we have seen before, this is a useful departure point. In the following, we would have a reality check with the domain experts that discusses whether the defined value ranges cover the domain well. While setting upper and lower limits is straight forward, assigning the similarity behaviour is not. Consecutively, we assume that numerical local similarity measures are distance functions and the question is how steep of a similarity decline should be chosen. We use polynomial functions to model similarity measure since they are more flexible and provide better convergence when using continuous numerical data. Therefore, we will focus on the polynomial function of the similarity measure and our goal is to determine their degree.

Taken this task in our application domain, we see an activity variation among different profiles, but also in the aggregation of activities over all profiles. We use box plots for visualizing the distributions and variations in our data set and transfer this into modelling local similarity measures.

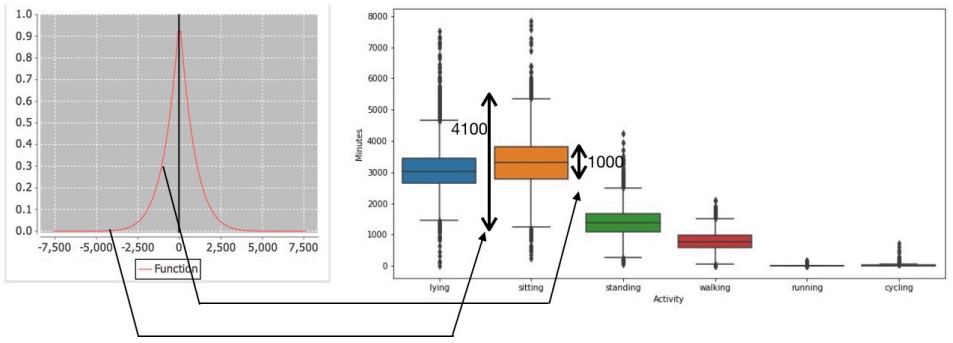


Figure 2: Example for Data-driven Local Similarity Modelling: On the left there is a screen shot of a polynomial similarity function for a value range between 0 and 7500. With the arrows we depict how the box-plot for sitting relates to the decrease of similarity at a certain distance. $IQR * 1.5$ method has been used for the box plots.

Fig 2 shows an example of a numerical local similarity measure. In the example, it is the total amount of sitting during six days. From there we look into the Q_1 and Q_3 which indicated the majority spread for the data set. We decided to take these values as reference points for determining the decrease of similarity.

Hence, creating a box-plot of the data set will allow modelling each activity attribute since we only take the Inter Quartile Range (IQR) and the range (min to max) into account:

$$\begin{aligned} r_1 &= IQR \\ r_2 &= range \end{aligned} \tag{1}$$

It represents the difference between upper (Q_3) and lower (Q_1) quartiles in the box-plot, that is $IQR = Q_3 - Q_1$.

We assume that all similarity functions are polynomial and adjust the polynomial degree of the similarity function such that

$$\begin{aligned} y(r_1) &\approx 0.30 \\ y(r_2) &\approx 0 \end{aligned} \tag{2}$$

We can observe in fig 2 how the similarity function varies after applying the methodology in equation 1 and 2. The bigger the polynomial degree, the steeper the similarity function and more precise the attribute values in retrieved cases. The decline in the similarity function is steeper in the beginning until at r_1 it reaches close to $y(r_1)$ and then decreases gradually until at r_2 it is approximately close to $y(r_2)$. This way, the similarity function covers the entire attribute range as well as the similarity measure range $[0, 1]$. While the choice of $y(r_1)$ and $y(r_2)$ depends on the domain-expert's knowledge and satisfaction with the outcome, we however experimented with different values and found these best suited for our application domain. We use this as the initial definition of similarity measures. If required, the function can of course be further customized if the relevant domain knowledge is available.

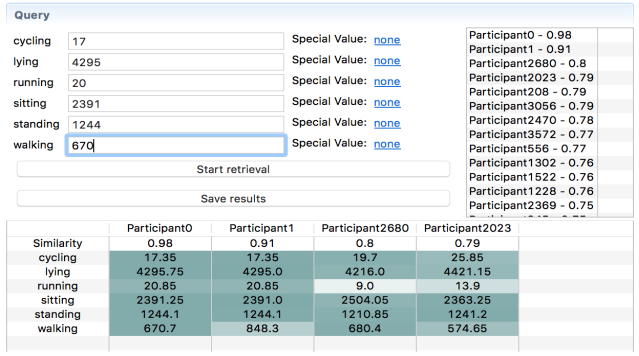


Figure 3: A Query and its retrieval result in the myCBR workbench

D.III Comparing Physical Activity Profiles

Once the casebase and similarity measures are in place, the model can be used to find similar profiles. Fig 3 shows the result of one such query retrieval in myCBR. The figure shows that the retrieved cases are sorted by similarity value in descending order, that is, most similar case are displayed at the top

while least similar are at the bottom. On the lower part of the screen shot the four most similar profiles are shown in a detailed view. The tool marks closer matches darker.

While the myCBR workbench indicates that we can do a similarity-based retrieval, it is hard to judge how the CBR system works with increasing casebase or changing similarity measures. In the next section we will investigate how the casebase size and different retrieval methods perform in our application domain.

E Evaluation of Increasing Casebase Sizes and Retrieval Methods

A performance evaluation of the CBR model has been conducted using holdout-repeat cross-validation in which 200 random cases were held out to be used for testing. Therewith for each run our casebase consisted of 4428 cases. A test set, comprising of ten randomly selected cases from the held out set of 200 cases, represents a single epoch in the experiments and performance is reported using Mean Relative Error (MRE) as a measure of precision. Each experiment is repeated five times and the results are averaged over all the epochs.

For each query instance q_i in the test set, the number of similar cases retrieved r from the casebase is 20. The relative error of each activity is then computed between q_i and r for one case at a time. The errors are averaged to obtain MRE of each activity for q_i. The process is repeated for every q_i in the test set, that is, for $i = [1, 10]$.

The MRE of the six activities are added to get the total relative error for each q_i. MRE is then calculated by averaging the relative errors for the entire queried test set.

The total relative error T for each queried instance is calculated as:

$$T = \sum_{i=1}^{6} MRE(A_i)$$

where A is the activity type as they were introduced in section D.I. MRE for the each test set is calculated as:

$$MRE = \frac{\sum_{i=1}^{10} T_i}{10}$$

The experiments in this evaluation are performed in two ways: First, by calculating the MRE of retrieved instances against each queried test instance with increasing casebase size. Second, by comparing the different results obtained using the CBR model and k-NN regressor model.

E.I Increasing Casebase Size

This experiment focuses on the variation observed in MRE with the increasing size of the casebase. The CBR model was implemented using myCBR, however the tool does not support batch queries, which was the need of the hour for conducting the experiments for our work. To overcome this limitation, we used a myCBR Rest API [8] for batch querying the casebase using POST calls and the implementation was done in Python (version 3.6.3).

In this experiment, a test set is passed as a query using POST call when the casebase initially has 500 instances. Subsequently, MRE for that test set is calculated. 500 cases are then added to the casebase and the process is repeated until the casebase consists of the entire data set. The experiment is repeated five times, each with a different random test set. The average MRE of all the epochs for the given casebase size is shown in Fig 4.

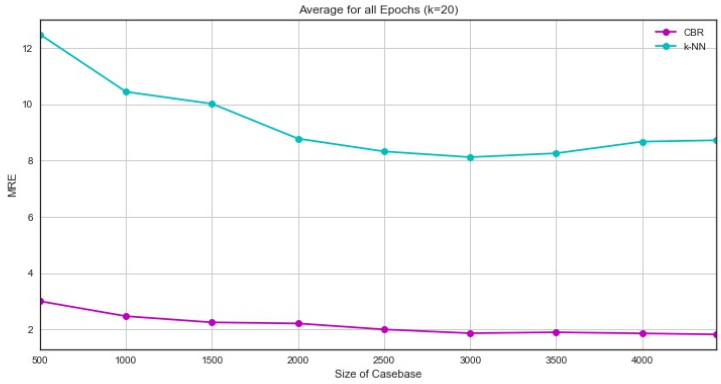


Figure 4: MRE comparison between the CBR model and k-NN regressor model with increasing casebase sizes (MRE is calculated for $k = 20$ retrieved cases)

[8] $https://github.com/kerstinbach/mycbr-rest-example$

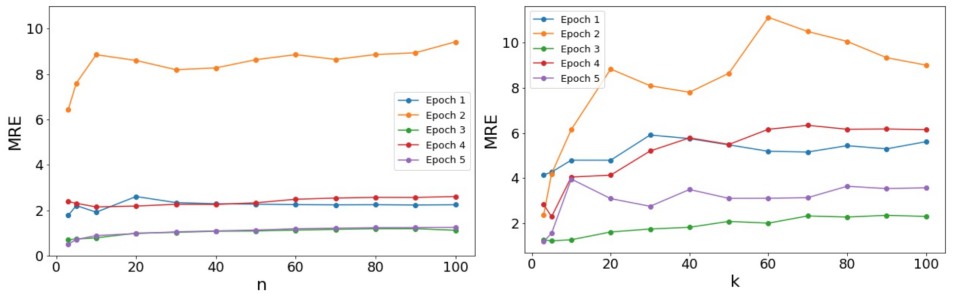
In order to have a comparison of the performance of the CBR model, the same experiment was conducted using k-NN regression model (with $k = 20$). The implementation of the k-NN regressor was done using Scikit learn (Pedregosa et al., 2011) library (version 0.19.1) in Python (version 3.6.3). The results obtained with the k-NN model are presented along with the results of the CBR model in Fig 4, where x-axis shows the size of the casebase (or size of data set for k-NN) and y-axis shows the MRE averaged over five epochs.

It can be observed from the results that MRE decreases steadily with increase in size of the casebase in the CBR implementation. However, the same cannot be said for k-NN, as the results show uncertain response to the increase in size of the data set. Even after introducing the entire data set, no improvement is observed. This decline in performance in k-NN is caused by the presence of outliers in the test set. CBR is able to estimate closest similar cases with respect to every activity for outliers very well, whereas k-NN cannot estimate the nearest neighbors with respect to every activity when presented with outliers. For instance, if there is an instance in the test set which has some or all attributes with values either below 25% or above 75% of the data range for those attributes in the data set, it leads to the k-NN algorithm computing nearest neighbors which are closer to the non-outlier attributes but farther from the outlier attributes. Thus, resulting in higher MRE even with an increased size of the data set.

E.II Selection of k

Selecting an appropriate value of k is crucial in determining the success or failure of a k-NN regressor model. To see how the error varies, we experimented with different values of k in the range [3,100]. Fig 5(b) shows the variation in MRE with the change in value of k. Here, x-axis shows the value of k and y-axis shows the MRE.

Although the determination of the closest similar profile in the CBR model is independent of n (number of retrieved cases), it is interesting to see how the MRE changes by varying n progressively. This allows us to further compare and contrast the performance of CBR model with that of k-NN model. Fig 5(a) shows the variation of MRE with increasing value of n in myCBR, where the x-axis shows the value of n and y-axis shows the MRE. It is clear from the results that the value of k in k-NN (refer Fig 5(b)) has a huge impact on the MRE for each epoch. The implication of this graph is that with an increase



(a) CBR (b) k-NN

Figure 5: Number of closest cases: On the left is the graph depicting the variation in MRE with the number of most similar cases retrieved (n) in CBR implementation. On the right is the graph for k-NN model depicting the variation in MRE with different values of k.

in k, more neighboring cases are taken into consideration which are either less similar altogether or less similar with respect to a subset of activities, resulting in the sudden variation in errors. Whereas the CBR model has a relatively smoother response in creating the number of retrieved similar cases. It can be argued from the results that lower values of k would have been more suitable due to less MRE. However, our aim in this work is not to predict using k-NN, but to find a number of nearest neighbors of the queried profile, which is why we chose $k = 20$ for our experiments. As our data set is large, $k = 20$ is reasonably acceptable for this application domain. Also, from CBR perspective, considering more neighboring profiles helps in making improvements to the similarity measure to a greater extent than considering just one neighbor profile.

E.III Composition of Error

As we are using activity data to find other similar profiles, it is important to know the error observed in the approximation of each activity in the similar profiles.

Fig 6 shows the MRE (in log) for each activity using both the approaches when introduced with the entire data set. The figure underlines that for inactive time (lying, sitting, standing) - which is the majority for the participants (see Table 2 and Fig 2) - the k-NN approach produces less of an

error. For moderate activities, like walking, both approaches are very close, while for rigorous activities, which we see only limited in the data set, the CBR approach produces much better results. This is very important for our overall aim of this work, as we eventually want to identify beneficial physical activity phenotypes.

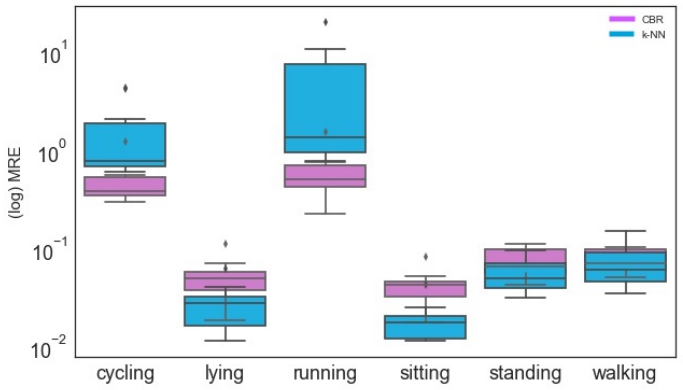
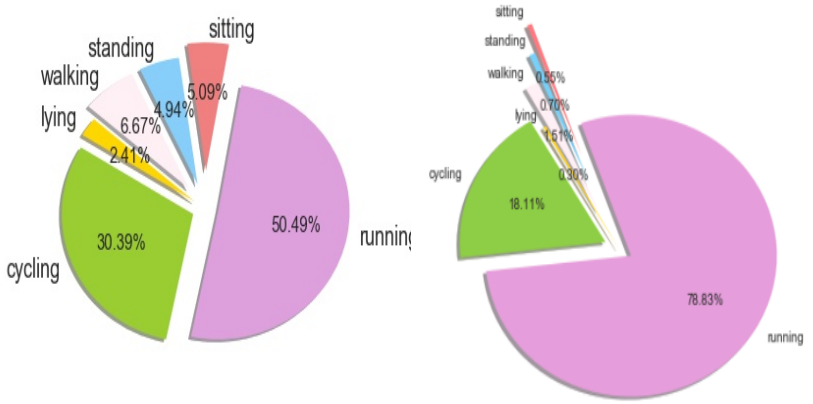


Figure 6: MRE per activity for the entire data set by the k-NN regressor and the CBR model

This observation is undermined by Fig 7, which shows the distribution of MRE for each of the activity calculated for both approaches after introducing the entire data set. It can be observed that in both k-NN and CBR, most of the error is attributed to the approximation of activity *running* (approx. 79% and 51% respectively). On the other hand, it is far lower in CBR, the result of which is relatively higher error composition of other activities as compared to those in k-NN. However, since these are compositional parts and convey only relative information, rather than concrete information, we must take into consideration the actual MRE, refer to Fig 4, which is significantly lower in case of CBR.

F Discussion

The experimental results shown in Fig 4 demonstrate that the CBR model performs well in finding similar physical activity profiles. While k-NN is



(a) CBR (b) k-NN

Figure 7: Error Composition for the CBR (a) and k-NN (b) model

able to well approximate four out of six physical activities when finding the nearest neighbours, however it fails miserably in finding with respect to the other two activities, which results in higher MRE. On the other hand, the CBR model is able to determine the most similar physical activity profiles with respect to every activity more closely, resulting in far lower MRE as compared to the k-NN model. Furthermore, k-NN is susceptible to outliers, which is the cause of increase in MRE even after introducing the entire data set. Whereas this is not an issue with the CBR model. In Fig 5 we observe very minor increase in MRE with increasing number of retrieved instances using CBR model, whereas the variations are more pronounced when using the k-NN model. These experiments demonstrate that the similarity modelling approach presented is working successfully for our application domain. Consequently, the CBR model significantly outperforms the k-NN algorithm and is more robust in finding similar physical activity profiles in a population. CBR approach can be applied to find and cluster similar activity groups, which will further be helpful in determining activity phenotypes.

G Conclusion and Future Work

In this paper, we presented an approach to model the local similarity measures for physical behaviour data in myCBR in a data-driven manner. This model can be applied on physical behaviour data acquired using wearable sensors to find, group and compare similar activity profiles. We have demonstrated through experiments and statistical evaluation how the CBR model outperforms the state-of-the-art k-NN regressor model. Thus, it can be concluded that CBR approach is a suitable and viable option for application such as this in the public health domain. It can further be utilized in determining activity phenotypes in order to provide personalized activity recommendations to participants and help slowly transform an inactive into a more active lifestyle. We have also demonstrated through experiments the effectiveness of similarity modelling approach presented in this paper for the public health domain and it will be safe to conclude that it can be transferred to other similar domains dealing with continuous numerical data.

The method presented can further be enhanced to automatically assign the local similarities based on the attributes' values in the casebase using machine learning techniques, similar to what (Gabel and Godehardt, 2015) presented in their paper. It can significantly reduce the efforts required to create new CBR models using different data sets from scratch.

In the future, we aim to extend our research towards compositional data analysis (Aitchison and Egozcue, 2005) on the HUNT4 data and applying CBR on the resulting compositional data. Compositional data analysis has been applied by researchers (Dumuid et al., 2017) for estimating the effect of change in physical activity behaviour for daily activities. Whether a change in one type of behaviour is beneficial or harmful for health depends on the compensatory shifts in other behaviours. The compositional nature of the HUNT4 data has therefore important consequences for both the analytical approach undertaken and interpretation of effects on health outcomes. Utilizing CBR for compositional data analysis will facilitate (i) getting insights into the behavioural characteristics between similar profiles in a population, (ii) understanding the association and co-dependency among various behaviours in different profiles, and (iii) identifying physical behaviour phenotypes.

Bibliography

Afshin A, Forouzanfar MH, and Reitsma MB. Health effects of overweight and obesity in 195 countries over 25 years. *New England Journal of Medicine*, 377(1):13–27, 2017. doi: 10.1056/NEJMoa1614362. PMID: 28604169.

Amira Abdel-Aziz, Marc Strickert, and Eyke Hüllermeier. Learning solution similarity in preference-based cbr. In Luc Lamontagne and Enric Plaza, editors, *Case-Based Reasoning Research and Development*, pages 17–31, Cham, 2014. Springer International Publishing. ISBN 978-3-319-11209-1.

J. Aitchison and J. J. Egozcue. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850, 2005. doi: 10.1007/s11004-005-7383-7.

Muhammad Arif and Ahmed Kattan. Physical activities monitoring using wearable acceleration sensors attached to the body. *PLOS ONE*, 10(7):1–16, 2015. doi: 10.1371/journal.pone.0130851.

Kerstin Bach and Klaus-Dieter Althoff. Developing Case-Based Reasoning Applications Using myCBR 3. In Ian Watson and Belen Diaz Agudo, editors, *Case-based Reasoning in Research and Development, Proceedings of the 20th International Conference on Case-Based Reasoning (ICCBR-12)*, pages 17–31. LNAI 6880, Springer, 2012.

Ralph Bergmann, Janet Kolodner, and Enric Plaza. Representation in case-based reasoning. *The Knowledge Engineering Review*, 20(03):209, 2005. doi: 10.1017/s0269888906000555.

Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46 (3):1–33, 2014. doi: 10.1145/2499621.

Boris Campillo-Gimenez, Wassim Jouini, Sahar Bayat, and Marc Cuggia. Improving case-based reasoning systems by combining k-nearest neighbour algorithm with logistic regression in the prediction of patients' registration on the renal transplant waiting list. *PLoS ONE*, 8(9), 2013. doi: 10.1371/journal.pone.0071991.

Luca Canensi, Giorgio Leonardi, Stefania Montani, and Paolo Terenziani. Multi-level interactive medical process mining. In Annette ten Teije, Christian Popow, John H. Holmes, and Lucia Sacchi, editors, *Artificial Intelligence in Medicine*, pages 256–260, Cham, 2017. Springer.

Dorothea Dumuid, Zeljko Pedisic, Tyman Everleigh Stanford, Josep-Antoni Martín-Fernández, Karel Hron, Carol A Maher, Lucy K Lewis, and Timothy Olds. The compositional isotemporal substitution model: A method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Statistical Methods in Medical Research*, 2017.

Thomas Gabel and Eicke Godehardt. Top-down induction of similarity measures using similarity clouds. In Eyke Hüllermeier and Mirjam Minor, editors, *Case-Based Reasoning Research and Development*, pages 149–164, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24586-7.

Shaker H. El-Sappagh and Mohammed Elmogy. Case representation and indexing. *Foundations of Soft Case-Based Reasoning*, page 34–74, 2004. doi: 10.1002/0471644676.ch2.

Erin K. Howie, Anne L. Smith, Joanne A. Mcveigh, and Leon M. Straker. Accelerometer-derived activity phenotypes in young adults: a latent class analysis. *International Journal of Behavioral Medicine*, 2018. doi: 10.1007/s12529-018-9721-4.

Eyke Hüllermeier and Patrice Schlegel. Preference-based cbr: First steps toward a methodological framework. In Ashwin Ram and Nirmalie Wiratunga, editors, *Case-Based Reasoning Research and Development*, pages 77–91, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-23291-6.

Aditya Khamparia and Babita Pandey. A novel method of case representation and retrieval in cbr for e-learning. *Education and Information Technologies*, 22(1):337–354, 2017. ISSN 1360-2357. doi: 10.1007/s10639-015-9447-8.

Harold W Kohl, Cora Lynn Craig, Estelle Victoria Lambert, Shigeru Inoue, Jasem Ramadan Alkandari, Grit Leetongin, and Sonja Kahlmeier. The pandemic of physical inactivity: global action for public health. *The Lancet*, 380(9838):294–305, 2012. doi: 10.1016/s0140-6736(12)60898-8.

J. Lagersted-Olsen, M. Korshøj, J. Skotte, I. Carneiro, K. Søgaard, and A. Holtermann. Comparison of objectively measured and self-reported time spent sitting. *International Journal of Sports Medicine*, 35(06):534–540, 2013. doi: 10.1055/s-0033-1358467.

I-Min Lee and Eric J Shiroma. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. *British Journal of Sports Medicine*, 48(3):197–201, 2013. doi: 10.1136/bjsports-2013-093154.

I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, and Peter T Katzmarzyk. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, 380(9838):219–229, 2012. doi: 10.1016/s0140-6736(12)61031-9.

Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, and et al. Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLOS Biology*, 15(1), 2017. doi: 10.1371/journal.pbio.2001402.

Michael Marschollek. A semi-quantitative method to denote generic physical activity phenotypes from long-term accelerometer data – the atlas index. *PLoS ONE*, 8(5), 2013. doi: 10.1371/journal.pone.0063522.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Kevin Plis, Razvan C. Bunescu, Cynthia R. Marling, Jay Shubrook, and Frank Schwartz. A machine learning approach to predicting blood glucose

levels for diabetes management. In *AAAI Workshop: Modern Artificial Intelligence for Health Analytics*, 2014.

Stephanie A Prince, Kristi B Adamo, Meghan Hamel, Jill Hardt, Sarah Connor Gorber, and Mark Tremblay. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 5(1):56, 2008. doi: 10.1186/1479-5868-5-56.

Olli T. Raitakan, Kimmo V K. Porkka, Simo Taimela, Risto Telama, Leena Räsänen, and Jorma S. Vllkari. Effects of persistent physical activity and inactivity on coronary risk factors in children and young adults the cardiovascular risk in young finns study. *American Journal of Epidemiology*, 140(3):195–205, 1994. doi: 10.1093/oxfordjournals.aje.a117239.

Michael M. Richter. The knowledge contained in similarity measures. In Manuela M. Veloso and Agnar Aamodt, editors, *Case-Based Reasoning Research and Development, Proc of the First International Conference, ICCBR-95*, volume 1010 of *LNCS*. Springer, 1995.

Sadiq Sani, Nirmalie Wiratunga, Stewart Massie, and Kay Cooper. knn sampling for personalised human activity recognition. In David W. Aha and Jean Lieber, editors, *Case-Based Reasoning Research and Development*, pages 330–344. Springer, 2017. ISBN 978-3-319-61030-6.

Barry Smyth and Pádraig Cunningham. Running with cases: A cbr approach to running your best marathon. In David W. Aha and Jean Lieber, editors, *Case-Based Reasoning Research and Development*, pages 360–374, Cham, 2017. Springer International Publishing. ISBN 978-3-319-61030-6.

Armin Stahl and Thomas R. Roth-Berghofer. Rapid prototyping of cbr applications with the open source tool mycbr. In *ECCBR '08: Proc. of the 9th European conference on Advances in Case-Based Reasoning*, pages 615–629, Berlin, 2008. Springer. ISBN 978-3-540-85501-9.

Mobyen Uddin and Amy Loutfi. Physical activity identification using supervised machine learning and based on pulse rate. *International Journal of Advanced Computer Science and Applications*, 4(7), 2013. doi: 10.14569/ijacsa.2013.040730.

Chi Pang Wen and Xifeng Wu. Stressing harms of physical inactivity to promote exercise. *The Lancet*, 380(9838):192–193, 2012. doi: 10.1016/s0140-6736(12)60954-4.

Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 uk biobank participants. *BioRxiv*, 2018. doi: 10.1101/187625.

Bangpeng Yao and Shao Li. Anmm4cbr: a case-based reasoning method for gene expression data classification. *Algorithms for Molecular Biology*, 5(1): 14, 2010. doi: 10.1186/1748-7188-5-14.

Similarity Measure Development for Case-Based Reasoning - A Data-Driven Approach

Deepika Verma, Kerstin Bach, Paul Jarle Mork

Abstract

In this paper, we demonstrate a data-driven methodology for modelling the local similarity measures of various attributes in a dataset. We analyse the spread in the numerical attributes and estimate their distribution using polynomial function to showcase an approach for deriving strong initial value ranges of numerical attributes and use a non-overlapping distribution for categorical attributes such that the entire similarity range [0,1] is utilized. We use an open source dataset for demonstrating modelling and development of the similarity measures and will present a case-based reasoning (CBR) system that can be used to search for the most relevant similar cases.

A Introduction

CBR has gained popularity in the recent years due to its novel approach to abstract and transfer domain-specific expert knowledge into a user-friendly tool which offers appropriate reasoning for solutions to problems ranging from simple daily life tasks to complex tasks which otherwise necessitate expert guidance.

Modelling the local similarities of attributes while preparing a CBR model can be a challenging task for small and simple, and large and complex data sets alike. In this paper, we direct our attention towards the knowledge engineering process of creating a CBR model and present a data-driven approach for modelling local similarity measures using the openly available User Knowledge Modelling dataset[1] in the myCBR workbench (Bach and Althoff, 2012; Stahl and Roth-Berghofer, 2008). The main contribution of this paper is a methodology for modelling the local similarity measures using a data-driven approach. We will showcase how the knowledge stored in a data set can be leveraged to define strong initial value ranges for both numerical and categorical attributes and therewith moderate and stratify the knowledge modelling process.

The remainder of this paper is organised into sections as follows: in section B, we discuss related work about the use of data-driven similarity measure development and its application in CBR, followed by section C wherein we present our similarity modelling approach. Finally, section D concludes the work presented in this paper.

[1] *https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling*

B Related Work

Similar to the preference-based similarity measure development framework presented by authors in Hüllermeier and Schlegel (2011); Abdel-Aziz et al. (2014), we are presenting a framework for modelling local similarity measures based on the data set available. Therewith we can tailor each similarity measure to the application domain. Using a data-driven approach for automatic similarity learning and feature weighting has been presented by Gabel and Godehardt (2015) where they trained a neural network to induce local and global similarity measures (Richter, 1995). While we are not automatically assigning the similarity measures, we use the existing cases to derive them.

C Data-driven Knowledge Modelling

In this section, we explain how we implement a CBR system that can be applied to find the most similar and relevant cases. We use the local-global-principle for tailoring the similarity measure for each attribute and thereby build a knowledge model (Richter, 1995). Once the local similarity measures are defined, we continue to use weighted sum for defining the global similarity.

Some of the most common challenges for utilizing any dataset for developing a CBR system are the identification of suitable dataset context for the problem at hand, definition of initial similarity measures, representation of cases and determination of valuable cases for populating the case base. In this section, we first describe how we populate the case base and generate cases in the developed case representation. Then we present our method for utilizing a given dataset to model the local similarity measures for both numerical as well as categorical attributes.

C.I Case Generation

Developing a case representation is the first step of the CBR system development. Depending on the domain and the available data this can be a challenging process on its own. For presenting our data-driven modelling technique, we use the User Knowledge Modelling dataset, which comprises of six attributes, five numerical and one categorical. The description of all the attributes is presented in table 1.

Attribute	Description
STG	The degree of study time for goal object materials
SCG	The degree of repetition number of user for goal object materials
STR	The degree of study time of user for related objects with goal object
LPR	The exam performance of user for related objects with goal object
PEG	The exam performance of user for goal objects
UNS	The knowledge level of user

Table 1: Description of attributes in User Knowledge Modelling dataset

The categorical attribute *USN* has four permitted values: *Very Low, Low, Middle, High*. Table 2 shows the data statistics of the numerical attributes in the dataset.

	STG	SCG	STR	LPR	PEG
count	403	403	403	403	403
mean	0.3531	0.3559	0.4576	0.4313	0.4563
min	0	0	0	0	0
max	0.99	0.90	0.95	0.99	0.99

Table 2: Data set Statistics

The case base is then populated by loading the dataset into the previously defined case representation in the myCBR workbench. A single case in myCBR is represented as shown in figure 1, where *User* is the name of the concept which comprises of six attributes present in the original dataset.

C.II Data-driven Similarity Measures Development

The local-global-principle requires both the local similarity measure on the attribute level and the global one on the conceptual to be defined.

Researchers in CBR domain face the challenge of balancing the input

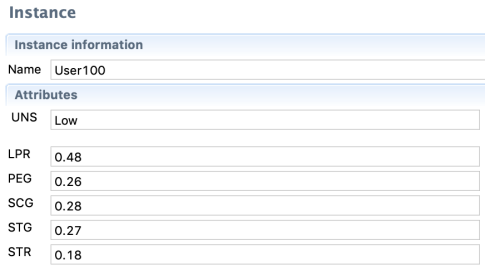


Figure 1: Case representation in myCBR

from the domain experts and the available data while modelling the local similarity measures for different attributes in myCBR. Having a criteria which can lead the knowledge modelling process is helpful for both parties. We therefore suggest to make use of the existing data in this process. While setting upper and lower limits for numerical attributes is straight-forward, assigning the similarity behaviour is not. Consecutively, we assume that local similarity measures for continuous numerical attributes are polynomial distance functions (due to their flexibility and better converging ability) and the question is how steep of a similarity decline should be chosen. Therefore, we focus on the polynomial function of the similarity measure for numerical attributes and our goal is to determine their degree. We use box plots for visualizing the distributions and variations in the data set and map this into modelling local similarity measures.

Figure 2 shows an example of a local similarity measure for a numerical attribute. From there we look into the Q_1 and Q_3, which indicate the majority spread of the attributes in the data set. In line with Abdel-Aziz et al. (2014); Verma et al. (2018), we decided to take these values as reference points for determining the decrease in similarity.

Hence, creating a box-plot of the data set will allow modelling each attribute since we only take the Inter Quartile Range (IQR) and the range (min to max) into account:

$$\begin{aligned} r_1 &= IQR \\ r_2 &= range \end{aligned} \tag{1}$$

It represents the difference between upper (Q_3) and lower (Q_1) quartiles in the box-plot, that is $IQR = Q_3 - Q_1$.

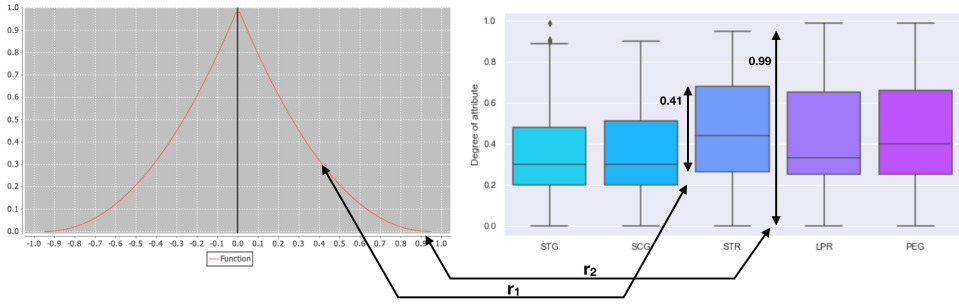


Figure 2: Example for Data-driven Local Similarity Modelling: On the left there is a screen shot of a polynomial similarity function for a value range between 0 and 1. With the arrows we depict how the box-plot for attribute *STR* relates to the decrease in similarity at a certain distance.

We assume that all similarity functions are polynomial and adjust the polynomial degree of the similarity function such that

$$\begin{aligned} y(r_1) &\approx 0.30 \\ y(r_2) &\approx 0 \end{aligned} \tag{2}$$

We can observe in figure 2 how the similarity function varies with respect to the attribute value after applying the methodology in equation 1 and 2. The bigger the polynomial degree, the steeper the similarity function and more precise the attribute values in retrieved cases. The decline in the similarity function is steeper in the beginning until at r_1 it reaches close to $y(r_1)$ and then decreases gradually until at r_2 it is approximately close to $y(r_2)$. This way, the similarity function covers the entire attribute range as well as the similarity measure range $[0,1]$. We use this as the initial definition of similarity measures.

While the local similarity measures for numerical attributes can be derived using their data distributions, assigning the similarity behaviour for categorical attributes can be challenging as it depends on whether or not there is a pre-existing relationship between the categorical values. In our dataset, the categorical attribute *UNS* has four permitted values which have an implicit relationship amongst each other. The local similarity measure for such an attribute can be modelled such that the relationship amongst the values is preserved while achieving the desired variation in the similarity

measure in the range [0,1], as shown in Figure 3. In case of no relationship amongst the values, the similarity of one value to every different value can be set to zero.

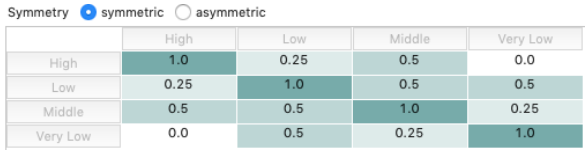


Figure 3: Similarity measure modelling for non-overlapping categorical attribute

C.III Retrieving Similar Cases

Once the casebase and similarity measures are in place, the model can be used to find similar cases. Figure 4 shows the result of one such query retrieval in myCBR. The retrieved cases are sorted by similarity value in descending order, that is, most similar case are displayed at the top while least similar are at the bottom. On the lower part of the figure, the four most similar *Users* are shown in a detailed view. The tool marks closer matches darker.

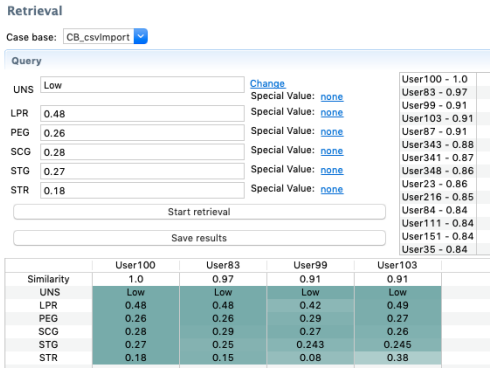


Figure 4: A Query and its retrieval result in the myCBR workbench

D Discussion and Conclusion

In this paper, we have presented an approach to model the local similarity measures of a given dataset in myCBR in a data-driven manner. Our approach can be applied on any dataset to model the similarity measures. A more detailed evaluation of our approach can be found in Verma et al. (2018) where we statistically evaluated its effectiveness using a public health domain dataset and showed that the CBR model created using our approach outperforms the k-NN regressor model in finding the most similar cases. The approach presented in this work can significantly reduce the efforts required to create new CBR models using different data sets from scratch. Therefore, it is safe to conclude that the approach works well on the used dataset and may also be applicable to other domains.

Bibliography

Amira Abdel-Aziz, Marc Strickert, and Eyke Hüllermeier. Learning solution similarity in preference-based cbr. In Luc Lamontagne and Enric Plaza, editors, *Case-Based Reasoning Research and Development*, pages 17–31, Cham, 2014. Springer International Publishing. ISBN 978-3-319-11209-1.

Kerstin Bach and Klaus-Dieter Althoff. Developing Case-Based Reasoning Applications Using myCBR 3. In Ian Watson and Belen Diaz Agudo, editors, *Case-based Reasoning in Research and Development, Proceedings of the 20th International Conference on Case-Based Reasoning (ICCBR-12)*, pages 17–31. LNAI 6880, Springer, 2012.

Thomas Gabel and Eicke Godehardt. Top-down induction of similarity measures using similarity clouds. In Eyke Hüllermeier and Mirjam Minor, editors, *Case-Based Reasoning Research and Development*, pages 149–164, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24586-7.

Eyke Hüllermeier and Patrice Schlegel. Preference-based cbr: First steps toward a methodological framework. In Ashwin Ram and Nirmalie Wiratunga, editors, *Case-Based Reasoning Research and Development*, pages 77–91, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-23291-6.

Michael M. Richter. The knowledge contained in similarity measures. In Manuela M. Veloso and Agnar Aamodt, editors, *Case-Based Reasoning Research and Development, Proc of the First International Conference, ICCBR-95*, volume 1010 of *LNCS*. Springer, 1995.

Armin Stahl and Thomas R. Roth-Berghofer. Rapid prototyping of cbr applications with the open source tool mycbr. In *ECCBR '08: Proc. of the 9th European conference on Advances in Case-Based Reasoning*, pages 615–629, Berlin, 2008. Springer. ISBN 978-3-540-85501-9.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Modelling similarity for comparing physical activity profiles - a data-driven approach. In Michael T. Cox, Peter Funk, and Shahina Begum, editors, *Case-Based Reasoning Research and Development*, pages 415–430, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01081-2.

Clustering of Physical Behaviour Profiles using Knowledge-intensive Similarity Measures

Deepika Verma, Kerstin Bach, Paul Jarle Mork

Abstract

In this paper, we reuse the Case-Based Reasoning model presented in our previous work (Verma et al., 2018) to create a new *knowledge-intensive similarity*-based clustering method that clusters a case base such that the intra-cluster similarity is maximized. In some domains such as recommender systems, the most similar case may not always be the desired one as a user would like to find the closest, yet significantly different cases. To increase the variety of returned cases, clustering a case base first, before the retrieval is executed increases the diversity of solutions. In this work we demonstrate a methodology to optimize the cluster coherence as well to determine the optimal number of clusters for a given case base. Finally, we present an evaluation of our clustering approach by comparing the results of the quality of clusters obtained using our knowledge-intensive similarity-based clustering approach against that of the state-of-the-art K-Means clustering method.

A Introduction

With the unprecedented growth in popularity of wearable activity trackers, acquiring reliable and objective physical behaviour data from users over a long period of time has become feasible. Activity trackers provide objectively measured basic activity statistics such as daily step count, miles run, heart rate among others while some selective trackers additionally provide activity recommendations to help user stay active throughout the day. While the validity and reliability of the activity trackers remains a topic of research (O'Driscoll et al., 2018), we conduct our research on the very premise of physical behaviour measured objectively, as opposed to self-reported (subjective) and that shall be the point of departure for our work ahead. Such objectively measured data present the opportunity to identify groups of people (or clusters) with similar physical behaviour (Marschollek, 2013; Howie et al., 2018). Further, this may provide a foundation for gaining new insights into the driving forces of physical behaviour in a population.

Clustering methods provide a simple yet powerful way to reveal underlying structure of the data and statistically understand the relationship between different data points. K-Means clustering is one of the most commonly employed state-of-the-art unsupervised machine learning method

for partitioning a given dataset into k clusters (MacQueen, 1967). Simple similarity metrics are used for calculating the similarity of the assigned cluster centroids to any given data point in the dataset in order to determine the cluster membership of each data point. The process repeats until no more changes in the position of centroids are observed. However, there are certain limitations to K-Means. It has a tendency to overlook data complexity (Yang et al., 2016) and moreover, is sensitive to outliers (Singh et al., 2011) and therefore can fail to give meaningful clusters in presence of many outliers in the dataset.

The challenge for most state-of-the-art clustering methods is the use of *knowledge poor similarity metrics* or simple distance metrics such as Hamming distance and Euclidean distance, among others. These metrics take into consideration only the syntactic difference between two data points, ignoring the coherence of each attribute or variable of a data point, thus leading to insufficient estimation of the similarity between them. In datasets where each variable takes on a value within a specific range elicits a requirement for modelling the local dependency for each variable. The similarity metric used must allow the existing knowledge to be brought to use for the assessment of similarity between data points in a dataset. Simple distance metrics can render the clusters incoherent in a complex dataset as opposed to cohesive clusters wherein the data points within a cluster are more similar to each other than to data points in another cluster. A solution to this problem can be formulated using Case-Based Reasoning (CBR) (Aamodt and Plaza, 1994), which employs a more knowledge-driven approach. Focusing on the semantic similarity between attributes rather than the syntactic similarity, the collective influence of each variable's importance on the final (global) similarity score will improve the clustering quality significantly by incorporating the existing knowledge in the dataset (Adam and Blockeel, 2015) and that CBR offers a more versatile approach to handle clustering of complex datasets (Müller and Bergmann, 2014).

In the sections that follow, we will use both *knowledge-intensive* as well as *knowledge-poor* similarity measures for cluster computation. We now hypothesize in this paper that using knowledge-intensive similarity measure as the metric for clustering the cases in a case base would create clusters wherein the cases within each cluster are semantically more similar to each other than to cases in the other clusters. The main contribution of this paper is a knowledge-intensive similarity based clustering method that can be used for any case base to compute clusters with high intra-cluster similarity. For

brevity sake, any mention of the term *similarity* from this point onwards shall be taken as a reference to the *knowledge-intensive similarity*, unless otherwise stated. The terms have also been used interchangeably.

This paper is organized into sections as follows: section B discusses the related work on similarity-based clustering, section C presents the application domain and elaborate on how similarity based clustering can be applied to identify clusters of physical behaviour profiles from the objective physical behaviour data; section D is dedicated towards our similarity based clustering algorithm; section E describes the dataset we use to test our algorithm; section F presents a set of experiments to evaluate our clustering approach, followed by section G discusses and conclude our work.

B Related Work

Application of clustering methods has played a major role in discovering the underlying patterns in public health data sets and understanding the characteristic differences among clusters. Identifying different clusters of similar physical behaviour patterns is similarly pivotal in understanding the physical activity characteristics of a population and will facilitate identification of different physical behaviour phenotypes[1]. Clustering has been previously applied by Marschollek (2013) on objectively measured physical behaviour data to identify four activity phenotypes using regularity, duration and intensity of activities as the pivotal attributes. Similar to their work, we aim at applying clustering, albeit knowledge-intensive similarity-based, on objectively measured physical behaviour data to identify phenotypes. Using a more probabilistic approach, Howie et al. (2018) identified five activity phenotypes for each gender using sex-specific latent class analysis. Although our approach differs from the one taken in their work, our long term goals and the target data are quite similar.

Similar to the self-efficacy based activity recommendation approach adopted by Baretta et al. (2019) to promote physical activity among adults, we aim to underpin activity recommendations based on the activity profile-assessed efficacy using a case-based approach in order to promote achievement of recommended physical activity goals[2]. A case-based

[1] *www.sciencedirect.com/topics/neuroscience/phenotype*

[2] *https : //www.who.int/ncds/prevention/physical − activity/guidelines − global − recommendations − for − health/en/*

marathon profile recommendation approach has been presented by Smyth and Cunningham (2017) to help marathon runners achieve their personal best. Using a different approach for improving the similarity-based retrievals in CBR, Müller and Bergmann (2014) presented a cluster-based indexing approach to make retrieval of most similar cases more efficient. While they use the similarity measure to construct a hierarchical cluster-tree which is used as an index for efficient retrieval, we use the similarity measure to create the clusters which can then be used as an index for retrieving relevant cases. Lucca et al. (2018) presented a framework for developing an index on clustered cases for improving query accuracy in agent simulation systems and making retrieval of relevant cases more efficient by organizing a large case base into smaller sub-case bases. Similarly, Cunningham (2009) introduced using similarity as a valid measure for selective sampling and generating solutions for unlabelled cases in clustered case bases.

Furthermore, Fanoiki et al. (2010) presented a cluster-based approach which facilitates the identification of relevant cases for a given query problem by considering the similarity relation among the cases within the case base with respect to their problem space as well the solution space. Their guiding principle being that the solutions of the most similar cases are likely to be similar if their problem descriptions are also similar. They formulate the solution by first selecting the cluster with the most similar problem description and then adapting the solution of the cases within that cluster. This is similar to what we intend to achieve for recommending activity goals.

C Clusters of Physical Behaviour Profiles

Real-time activity tracking and systematic physical activity recommendations remind users to help them stay active throughout the day. This is especially useful for sedentary individuals (Lagersted-Olsen et al., 2013). Prolonged uninterrupted bouts of sedentary behaviour are known to be detrimental to health (Saunders et al., 2012). In addition to the type of physical activity, the intensity of the moderate to vigorous activity performed also has an impact on the overall health outcomes (Ekelund et al., 2019).

The importance of enough sedentary behaviour has also been acknowledged since both high as well as low ends of the activity spectra are necessary in the right balance in order to promote good health (Coenen et al., 2018). However, the existing state-of-the-art trackers provide

approximately the same recommendations with slight variation to every user. Recommending activity goals to an individual which are challenging, yet achievable is more beneficial for improving their health as opposed to recommending either unachievable or not challenging enough activity goals (Baretta et al., 2019). Using an example from our dataset, we demonstrate how a CBR system can be used to identify unique clusters of physical behaviour profiles and how evidence-based experience of other similar profiles can be used to underpin activity recommendations for an individual.

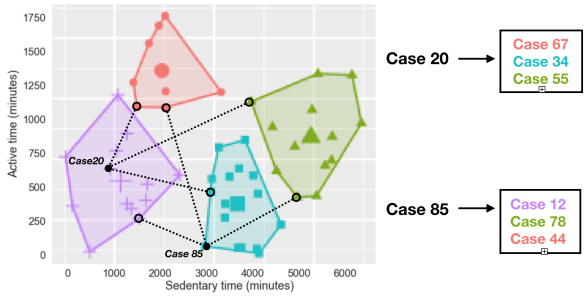


Figure 1: Example: A potential set of similarity-based clusters and how they can be utilised to recommend achievable activity goals to a user. The x-axis and y-axis show total sedentary duration and active duration (in minutes), respectively, over a period of six days.

Suppose we identify four clusters of physical behaviour profiles, as shown in figure 1 (we use a small subset of the original dataset for clarity in the visualization), in our dataset (see section E. The aim is to provide a user a diverse set of adapted most similar profiles from other clusters as recommendations, ranked by their similarity (such that lower similarity indicates more challenging goal). For instance, to recommend activity goals to *case 20*, the system can select one most similar case from each cluster other than its' member cluster and return the set of adapted profiles ranked by similarity to offer a diverse set of options for the user to choose their goal from. The most similar profile, *case 67* appears to be a challenging as well as an achievable goal for *case 20*. Therefore, it might be advisable for *case 20* to try and get closer to the adapted activity profile of *case 67* if they wish to challenge themselves while at the same time achieve the recommended activity goals. Similarly for *case 85*, *case 12* appears to be a challenging and achievable goal. Therefore, in this case, it might be advisable for *case 85* to

try and get closer to the adapted activity profile of *case 12* in order to become more physically active.

Large and complex datasets such as the objective measurements for the HUNT4[3] (see section E) study require pre-processing and organization of the case base to improve the overall performance of a CBR system. We address this topic by identifying unique clusters of different physical behaviour within the HUNT4 dataset using our similarity-based clustering method. We direct our attention solely towards understanding the behavioural characteristics of a sample population that contribute to differences in physical activity and sedentary behaviour which could allow for designing improved recommendations tailored to each phenotype for an innovative, yet effective active lifestyle management intervention. To elicit greater improvements in the existing infrastructure of activity recommendations, radical shift in the use and application of the existing methodologies may be required.

D Knowledge-intensive Similarity-based Clustering

Unsupervised machine learning methods provide a way of inferring underlying patterns or structure in a given dataset without any reference to known outcomes and therefore, is a viable option for our problem. We have a dataset consisting of 9034 physical behaviour profiles and look for clusters that represent meaningful physical behaviour types. Each cluster should be semantically coherent. While the state-of-the-art clustering methods such as K-means do provide a set of clusters, the profiles within each cluster are not guaranteed to be very semantically similar to each other since these methods use *knowledge-poor similarity measures* or simple distance measures.

As we have shown in our previous work (Verma et al., 2018), CBR outperforms the k-NN method in finding the most similar physical behaviour profiles. We therefore use the similarity score as the measure for clustering the profiles in our dataset. Our approach for using similarity as the metric for clustering extends the conventional approach of similarity in CBR by allowing to model and further utilize the similarity measures which are aligned with domain expert knowledge. Algorithm 1 introduces the knowledge-intensive

[3] *https://www.ntnu.no/hunt4*

similarity-based clustering algorithm used in our work.

Input : case base C, number of clusters n
Output: n clusters
initialization: assign n random cases as centroids-$\{c_n\}$
Determine Cluster Membership
for *each case k in C* **do**
  compute $sim(k,c_j)$, $\forall j \in 1,...,n$
  assign k to most similar centroid
end
Update Cluster Centroids
for *each* c_j *in* $\{c_n\}$ **do**
  compute $meanSim_j = \frac{1}{|S_j|} \sum_{\forall k_i \in S_j} sim(k_i, c_j)$
  find case m in S_j such that
   $sim(m,c_j) \approx meanSim_j$
  assign m as the new centroid c_j
end
Repeat until centroids converge
S_j denotes the set of cases in cluster c_j.

Algorithm 1: Knowledge-intensive Similarity-based Clustering Algorithm

The algorithm initially assigns n cases as centroids at random and then computes the clusters using the similarity score of each case to each centroid. As the similarity-based clustering method operates on the similarity score between each case and each centroid to determine its' cluster membership, it is independent of the data type. As a result, one advantage of this method is that it can be applied to different types of data sets other than just numerical, for example categorical or mixed datasets, which otherwise proves to be challenging when using the conventional clustering methods. Once the similarity measures are in place, the user is freed from the trouble of taking care of the data types before applying this *knowledge-intensive similarity*-based clustering method.

E Dataset

The data set used in this work is the objectively measured physical activity data collected during the fourth round of the HUNT[4] cohort study. The data

[4]https://www.ntnu.no/hunt

collection in HUNT4 spanned over 18 months and was finished in February 2019. Each person who volunteered to participate in the objective physical activity data collection was fitted with two tri-axial accelerometers, AX3 Axivity[5], one on the lower back and another on the thigh and wore them for a period of seven consecutive days. Objective measurements of a total of 35449 participants have been collected and basic physical activities have been assigned (see Table 1).

Table 1: Activity Descriptions.

Activity	Description
Lying	The person is lying down
Sitting	When the person's buttocks is on the seat of a chair or something similar
Standing	Upright, feet supporting the person's body weight
Walking	Locomotion towards a destination with one or more strides
Running	Locomotion towards a destination, with at least two steps where both feet leave the ground during each stride
Cycling	The person is riding a bicycle

Before populating the CBR system, we pre-process the data to obtain the same amount of data per participant. Therefore we decided to only include participants who have full six days of measured data. Furthermore, we remove any record containing zero minutes for *lying, standing, sitting* and less than one minute for *walking* activity as well as records where the sum of all activities exceeds 1440 minutes for a day (which represents the total minutes in a day). Due to various reasons including discomfort, sensor failure, loss or removal of sensor, our dataset reduced to 31113 participants, out of which we randomly sample 9034 participants while maintaining the overall distribution of activities for our experimental evaluation. Figure 2 shows the distribution of the six activities in the dataset.

F Experimental Evaluation

We implemented the *knowledge-intensive similarity*-based clustering algorithm in Java (version 1.8) using the java implementation of myCBR tool[6]. The CBR model for our dataset has been created in the myCBR

[5] $https://axivity.com/downloads/ax3$

[6] $https://github.com/ntnu-ai-lab/mycbr-sdk$

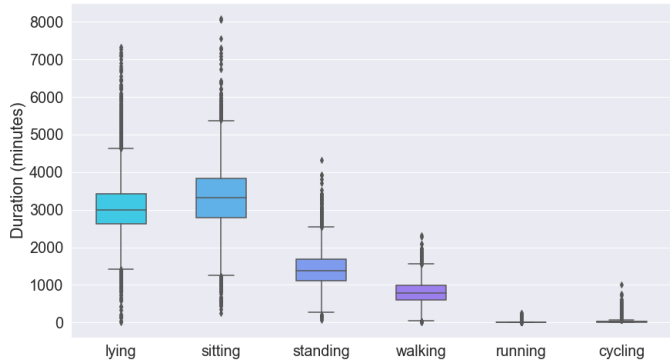


Figure 2: Summary of test dataset (9034 participants): Distribution of minutes spent per activity category over a period of six days

workbench (Stahl and Roth-Berghofer, 2008) by importing the data from a csv file. Similarity modelling of each activity attribute has been carried out in the same data-driven manner as we have presented in our last work (Verma et al., 2018). We then used the CBR model in our java implementation of the algorithm to compute any desired number of clusters.

F.I Coherent Clusters

A new set of centroids in the *knowledge-intensive similarity*-based clustering algorithm may or may not give better mean similarity of clusters than the previous centroids. We can observe in figure 3, the mean similarity of clusters varies to a large degree with each progressive round of clustering, wherein each round represents a new set of centroids. These variations occur due to change in cluster membership of the cases. As the membership of cases in the case base evolves over several rounds, the movement of cases, especially the *edge* cases from one cluster to another may result in increase in the mean similarity of the exiting cluster and decrease in that of the joining cluster or vice-versa, thereby introducing positive as well as negative variations in the cluster mean similarity. These variations make it challenging to determine the optimal centroids and clusters at any given point in the algorithm.

Direct optimization of similarity-based clustering is an NP-hard problem

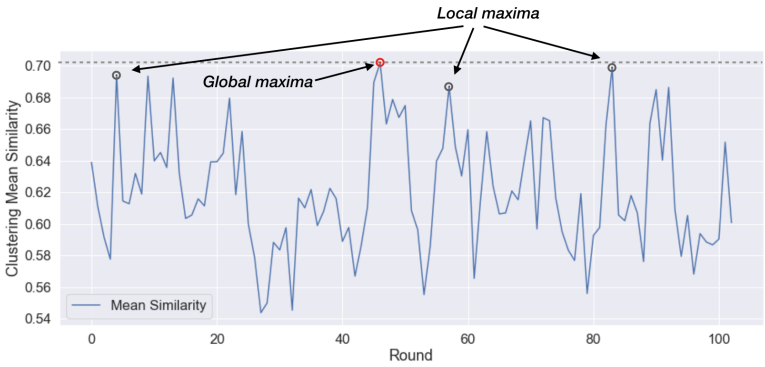


Figure 3: Variation in average similarity of ten clusters over multiple rounds

(Yang et al., 2016). To overcome this challenge, we employ a strategy where the algorithm looks s steps or rounds to the future to check if it finds a set of centroids with a higher mean similarity than the current set of centroids. It declares convergence only when it does not find any new set of centroids with a higher mean similarity than the current maximum mean similarity after s steps. The objective now is to determine the step size s. It can be observed in figure 3 that the mean similarity undergoes considerable amount of variation over multiple rounds. Therefore, s must be set large enough to foresee enough number of rounds before declaring convergence, but small enough to be computationally inexpensive for large datasets. The hypothesis here is that the probability of falling into a local maxima is less if the step size s is large enough to accommodate the variation observed in the mean similarity of clusters over multiple rounds, wherein each round consists of a new set of centroids.

We can observe in the figure 4, with the increase in the number of clusters, there is a decrease in the difference between the mean similarity achieved at any given s and the maximum mean similarity. This indicates an inverse relation between step size s and the number of clusters n. The value of s may differ depending on the size of the dataset and the number of clusters chosen, however, for our dataset, $s = 50$ seems to give a fair trade-off between time complexity and cluster coherency.

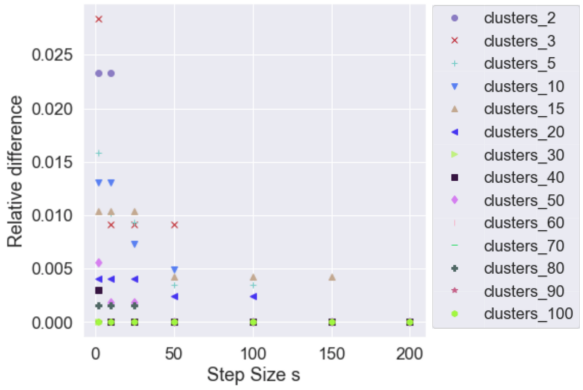


Figure 4: Relative difference in the mean similarity of clusters with the step size s for number of clusters n in the range [2,100]: The y-axis of the graph represents the difference between the maximum mean similarity and mean similarity achieved at s, displayed by the x-axis, for each n

F.II Number of Clusters

Clustering allows you to split a given data set into clusters according to a similarity metric, but one must specify the desired number of clusters in advance. Determining the optimal number of clusters in unsupervised clustering is a fundamental challenge and can be a daunting task. One way to determine the optimal number of clusters in K-Means is the elbow method, which involves plotting the sum of squared errors (SSE) against the number of clusters. As SSE decreases with the increase in number of clusters, the optimal number of clusters is observed by noting the *elbow* in the graph. In our case however, as we are operating on the mean similarity of clustering which is expected to increase with the increase in the number of clusters, we will have a reverse elbow graph.

To determine the optimal number of similarity clusters we plot the mean similarity of clusters against the number of clusters. With $s = 50$, we computed n clusters in the range [2,100] in order to learn the optimal number for our dataset. Five epochs were computed with n randomly chosen cases as initial centroids, wherein each epoch consists of reassignment of cases and recomputing the centroids until the clusters converge. Afterwards, an average was computed from the mean similarity values of all the five epochs. The results are shown in figure 5, where it can be observed that the

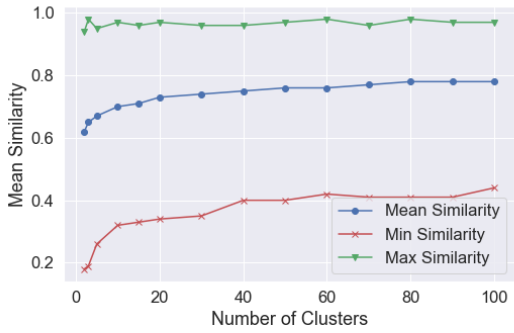


Figure 5: Similarity within clusters for the knowledge-intensive, similarity-based clustering(step size s = 50)

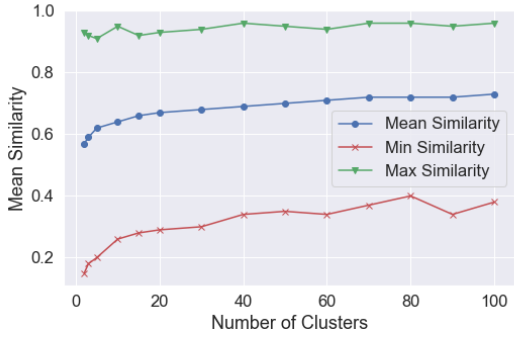


Figure 6: Similarity within clusters for the K-Means clustering method

mean similarity increases gradually until 20 clusters, followed by a slow but steady increase. This indicates the maximum optimal number of clusters for our dataset is 20 or less. We need a more detailed analysis in order to uniquely identify the different phenotype clusters from our dataset and aim at achieving this goal using similarity-based clustering.

F.III Assessment of Cluster Quality

We now evaluate the quality of the computed clusters within our dataset using our similarity-based clustering approach. We present an evaluation by comparing the performance of the proposed similarity-based clustering

method with that of state-of-the-art K-Means clustering method. The implementation of K-Means clustering algorithm was done using Scikit Learn library in Python (version 3.6.3) (Pedregosa et al., 2011) .

For comparing the results for both methods, we needed a common metric to base the comparison on. Since our aim is to have clusters with high degree of intra-cluster similarity, we decided to take the mean, minimum and maximum similarity as the metric for comparing the methods. However, K-Means does not compute semantic similarity between two given data points. To overcome this hurdle, we implemented a Rest API function in the myCBR java package which allows us to compute the similarity of any two given cases, provided that the attribute values are within their respective range as defined in the CBR model. We then used POST calls to calculate the similarity between each case and its cluster centroid for each cluster obtained using K-Means implementation. Five epochs were computed for both K-Means and similarity-based clustering methods. Each epoch consisted of reassignment of cases and recomputing the means until the clusters converge. An average was then computed of all the five epochs. The number of clusters n computed in each epoch were in the range [2,100].

Figures 5 and 6 show the minimum and maximum similarity for all the clusters in addition to the mean similarity for both similarity-based clustering and K-Means clustering. It can be observed from the results that the mean similarity and the minimum similarity for each number of clusters n are higher in similarity-based clustering, however there is not much difference in the maximum similarity. To further verify the difference in the results obtained by our algorithm and K-Means, we performed a t-test at significance level $\alpha = 0.01$ and $\alpha = 0.05$ for the mean similarity values of the clusters obtained using both the methods. The result is: $t\text{-}value = 2.87$, $p\text{-}value = 0.008$; which is significant at both α.

Although the measurable difference between results obtained using K-Means and similarity-based clustering appears to be small, the t-test results show that the results obtained are significantly different. Moreover, the difference lies in the quality of the clusters obtained using both the methods. As stated previously, our objective in this work has been to create clusters wherein the cases within each cluster are more similar to each other than to cases in other clusters. In other words, if we were to query for m similar cases for a particular case, say Participant 8921, we would expect the most similar cases to be in the same cluster as the queried case rather than in some other cluster(s), except perhaps for the edge cases. We can examine

this by querying the case base in the myCBR workbench and then verifying the cluster labels of the m most similar cases in the clusters obtained using both K-Means and similarity-based clustering methods. We choose $n = 20$ and make retrievals using two randomly chosen cases with $m = 6$. Figure 7 presents the results.

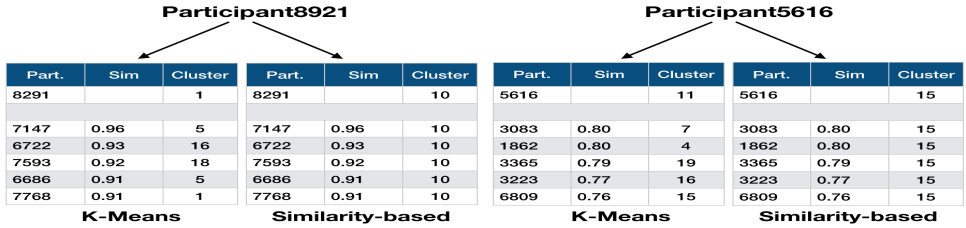


Figure 7: Examples showing the quality of clusters for k-Means vs similarity-based clusters. [Part.: Participant, Sim: Similarity]

Taking as reference the top most record, which is the queried case itself, we can now compare and contrast the difference in the quality of the clusters obtained using both the methods. In both the examples, the most similar cases in the similarity-based clusters are placed in the same cluster. On the other hand, most of the similar cases are placed in different clusters in the K-Means clusters. The examples presented in figure 7 support our hypothesis that the quality of clusters achieved using our approach is much superior.

G Discussion and Conclusion

In this paper, we have presented a clustering algorithm which uses *knowledge-intensive similarity* as the metric for computing clusters in a case base. We presented an evaluation using the clustering method in a CBR application built for the HUNT4 physical behaviour dataset. The method computes clusters and demonstrates how coherent clusters can be obtained using an optimization strategy (see section F.I). The experimental results shown in figures 5 and 6 along with the examples presented in figure 1 inevitably demonstrate the coherence as well as the diversity of the clusters obtained using our similarity-based clustering approach.

As stated previously, the conventional clustering methods such as K-Means have certain limitations which can be overcome using CBR. K-Means

tends to overlook the complexity of the data and puts emphasis on the attributes which have a dominant presence in the data (such as *lying*) while ignoring the smaller (such as *running*) but significant attributes. While a small-scale change in the small attributes may not result in a very large difference in the similarity score, it can however change the order of the similar cases. And thus, even though the cases in each K-Means cluster have a fairly high similarity to their cluster centroid, they are not necessarily very similar to each other.

We have demonstrated experimentally the clusters obtained using our similarity-based clustering approach have higher intra-cluster similarity amongst the cases as opposed to the clusters obtained using the state-of-the-art K-Means clustering method. The difference in the results obtained has been found to be statistically significant. Therefore, it is safe to conclude that our hypothesis is correct and the proposed similarity-based clustering algorithm provides better clusters than the K-Means clustering method. The proposed algorithm is a suitable and viable option for our application and gives the desired coherent clusters. The proposed similarity-based clustering method can nevertheless be applied to other datasets as well, including mixed datasets since the method is independent of the data types.

In future, we will investigate the physical behaviour profiles in more detail and use sequential physical behaviour data for clustering profiles by adding on information such as the intensity, frequency and duration of the activity bouts. The guidelines on physical activity make it evident that there is a necessity to develop recommendations that address the links amongst the type, duration, intensity, frequency and the total amount of physical activity necessary to be done by an individual in order to prevent non-communicable diseases and general health issues. We will extend our work to address this challenge by using similarity-based clustering to determine more specialized clusters and attempt to steer towards identifying the physical behaviour phenotypes in our dataset.

Bibliography

Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, 7(1), 1994.

A Adam and H Blockeel. Dealing with overlapping clustering: A constraint-based approach to algorithm selection. *CEUR Workshop Proceedings*, 1455, 01 2015.

Dario Baretta, Fabio Sartori, Andrea Greco, Marco D'Addario, Riccardo Melen, and Patrizia Steca. Improving physical activity mhealth interventions: Development of a computational model of self-efficacy theory to define adaptive goals for exercise promotion. *Advances in Human-Computer Interaction*, 2019, 2019.

Pieter Coenen, Lisa Willenberg, Sharon Parry, Joyce W Shi, Lorena Romero, Diana M Blackwood, Christopher G Maher, Genevieve N Healy, David W Dunstan, and Leon M Straker. Associations of occupational standing with musculoskeletal symptoms: a systematic review with meta-analysis. *British Journal of Sports Medicine*, 52(3), 2018.

Padraig Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Trans. Knowl. Data Eng.*, 21, 11 2009.

Ulf Ekelund, Wendy J Brown, Jostein Steene-Johannessen, Morten Wang Fagerland, Neville Owen, Kenneth E Powell, Adrian E Bauman, and I-Min Lee. A systematic review and harmonised meta-analysis of data from 850 060 participants. *British Journal of Sports Medicine*, 53(14), 2019.

Titilola O. Fanoiki, Isabela Drummond, and Sandra A. Sandri. Case-based reasoning retrieval and reuse using case resemblance hypergraphs. In *International Conference on Fuzzy Systems*, 2010.

Erin K. Howie, Anne L. Smith, Joanne A. McVeigh, and Leon M. Straker. Accelerometer-derived activity phenotypes in young adults: a latent class analysis. *International Journal of Behavioral Medicine*, 25(5), Oct 2018. ISSN 1532-7558.

Julie Lagersted-Olsen, Mette Korshøj, Jörgen Skotte, Iara Gabriel Carneiro, Karen Søgaard, and Andreas Holtermann. Comparison of objectively measured and self-reported time spent sitting. *International journal of sports medicine*, 35 6, 2013.

Marcos R. B. Lucca, Alcides G. Lopes Junior, Edison Pignaton de Freitas, and Luis A. L. Silva. A case-based reasoning and clustering framework for the development of intelligent agents in simulation systems. In *FLAIRS, Florida*, 2018.

J. MacQueen. *Some methods for classification and analysis of multivariate observations*. University of California Press, Berkeley, Calif., 1967.

Michael Marschollek. A semi-quantitative method to denote generic physical activity phenotypes from long-term accelerometer data—the atlas index. *PLOS ONE*, 8(5), 05 2013.

Gilbert Müller and Ralph Bergmann. A cluster-based approach to improve similarity-based retrieval for process-oriented case-based reasoning. In *Frontiers in Artificial Intelligence and Applications*, ECAI'14. IOS Press, 2014. ISBN 978-1-61499-418-3.

Ruairi O'Driscoll, Jake Turicchi, Kristine Beaulieu, Sarah Scott, Jamie Matu, Kevin Deighton, Graham Finlayson, and James Stubbs. How well do activity monitors estimate energy expenditure? a systematic review and meta-analysis of the validity of current technologies. *British journal of sports medicine*, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2011.

Travis J Saunders, Richard Larouche, Rachel C Colley, and Mark S Tremblay. Acute sedentary behaviour and markers of cardiometabolic

risk: a systematic review of intervention studies. *Journal of nutrition and metabolism*, 2012.

Kehar Singh, Dimple Malik, and Naveen Sharma. Evolving limitations in k-means algorithm in data mining and their removal. *International Journal of Computational Engineering and Management*, 12, 2011.

Barry Smyth and Pádraig Cunningham. Running with cases: A cbr approach to running your best marathon. In David W. Aha and Jean Lieber, editors, *CBR Research and Development*, Cham, 2017. Springer.

Armin Stahl and Thomas R. Roth-Berghofer. Rapid prototyping of cbr applications with the open source tool mycbr. In *ECCBR '08*. Springer-Verlag, 2008. ISBN 978-3-540-85501-9.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Modelling similarity for comparing physical activity profiles - a data-driven approach. In Michael T. Cox, Peter Funk, and Shahina Begum, editors, *CBR Research and Development*, Cham, 2018. Springer. ISBN 978-3-030-01081-2.

Zhirong Yang, Jukka Cor, er, and Erkki Oja. Low-rank doubly stochastic matrix decomposition for cluster analysis. *Journal of Machine Learning Research*, 17(187), 2016.

Exploratory Application of Machine Learning Methods on Patient Reported Data in the Development of Supervised Models for Predicting Outcomes

Deepika Verma, Duncan Jansen, Kerstin Bach, Mannes Poel, Paul Jarle Mork, Wendy Oude Nijeweme d'Hollosy

Abstract

Patient-reported outcome measurements (PROMs) are commonly used in clinical practice to support clinical decision making. However, few studies have investigated machine learning methods for predicting PROMs outcomes and thereby support clinical decision making. Using two datasets consisting of PROMs from 1) care-seeking low back pain patients in primary care who participated in a randomized controlled trial, and 2) patients with neck and/or low back pain referred to multidisciplinary biopsychosocial rehabilitation, we investigate data science methods for data prepossessing and evaluate selected regression and classification methods for predicting patient outcomes. The results show that there is a potential for machine learning to predict and classify PROMs. The prediction models based on baseline measurements perform well, and the number of predictors can be reduced, which is an advantage for implementation in decision support scenarios. The classification task shows that the dataset does not contain all necessary predictors for the care type classification. Overall, the work presents generalizable machine learning pipelines that can be adapted to other PROMs datasets. This study demonstrates the potential of PROMs in predicting short-term patient outcomes. Our results indicate that machine learning methods can be used to exploit the predictive value of PROMs and thereby support clinical decision making, given that the PROMs hold enough predictive power

A Introduction

While the application of machine learning (ML) methods is expanding into new clinical areas, both in medical research and clinical practice (Jiang et al., 2017; Yu et al., 2018), these methods have rarely been used on patient-reported outcome measurements (PROMs). PROMs are used commonly for health conditions that are difficult to assess with objective measurements, such as non-specific musculoskeletal pain and mental disorders. The predictive capabilities of ML methods, combined with clinical expertise, may increase the precision of clinical decision-making and thereby improve patient outcomes in these conditions (Holmes et al., 2017). To the best of our knowledge, no prognostic models based on ML methods are currently in clinical use for predicting outcomes among patients with non-specific

musculoskeletal conditions, such as neck pain and low back pain. These conditions are among the leading causes of disability worldwide (Hurwitz et al., 2018) and improving the precision of clinical decision-making to improve patient outcomes will likely have a substantial impact on their disability burden.

Predicting outcomes from PROMs in patients with neck and/or low back pain (NLBP) is a challenging task owing to the subjective nature of the data. Nevertheless, some recent studies have shown promising results in applying ML methods. In a study by Nijeweme-d'Hollosy et al. (2018), binary classification models trained on PROMs data were used to predict whether low back pain patients should be referred to a Multidisciplinary biopsychosocial rehabilitation (MBR) program or undergo surgery. The authors concluded that the ML models show small to medium learning effects. Another study showed that a ML least shrinkage selection operator approach performs well in predicting pain-related disability at 2-year follow-up among older adults with NLBP (Fontana et al., 2019).

The current study continues this line of research, intending to investigate to what extent different ML methods applied to PROMs data can identify predictors of outcomes and predict outcomes among patients with non-specific NLBP. The research question addressed in this work is: *Can Machine Learning methods make predictions using patient-reported data to facilitate the shared decision-making process for patients with NLBP?*.

B Background

Early and thorough assessment of non-specific low back pain is recommended to support a clinician's treatment planning for patients at increased risk of poor outcome (Lin et al., 2020). MBR is a commonly used treatment approach that targets biological, psychological, and social influences on low back pain (Saragiotto et al., 2015). However, this treatment approach is costly and time-consuming and the decision on whether a patient should start an MBR program is challenging. Supported self-management via web or mobile application is another alternative treatment approach that has gained popularity in recent years (Machado et al., 2016). One such decision support system (DSS) delivered via mobile application has been implemented in the selfBACK project (Mork and Bach, 2018). selfBACK DSS was developed to facilitate, improve, and reinforce self-management of non-specific LBP.

The core idea is to empower patients to take control of their symptoms and treatment.

PROMs are a valuable source of information but few studies have exploited PROMs in the context of applying ML methods. Rahman et al. (2018) performed a study, aimed at predicting pain volatility among users of a supported self-management delivered via a mobile application ("Manage My Pain"). Unsupervised ML methods were used to cluster the users followed by supervised ML methods to predict pain volatility levels at 6-month follow-up using in-app PROMs (130 in total). The best accuracy was 70%, achieved using Random Forest. In a follow-up study, Rahman et al. (2019) addressed the topic of identifying the most important predictors of pain volatility using different feature selection methods and found that similar prediction accuracy (68%) can be achieved using only a few predictors (9 features). In another study, Harris et al. (2019) compared the performance of four supervised ML models including Logistic, LASSO, Gradient Boosting Machines, and Quadratic Discriminant Analysis for predicting whether or not a patient achieves a minimal clinically important difference (MCID) in several pain and function related outcomes at 1-year post knee arthroplasty. Using preoperative PROMs as predictors, they found that similar performance can be achieved across different models for various outcomes by varying the number of inputs. None of the models was found to be superior for all the outcomes. In contrast, Fontana et al. (2019) found that LASSO performs better than Gradient Boosting Machines and Support Vector Machines in predicting MCID at 2-year follow-up among patients undergoing knee or hip arthroplasty. Similarly, Huber et al. (2019) compared the performance of eight supervised ML models for predicting MCID at six months among patients undergoing knee or hip replacement. Preoperative PROMs were used as predictors, and the results showed that Gradient Boosting machines yielded the most accurate prediction.

C Datasets

In this section we describe the two datasets used in this work to build classification and regression models for PROMs.

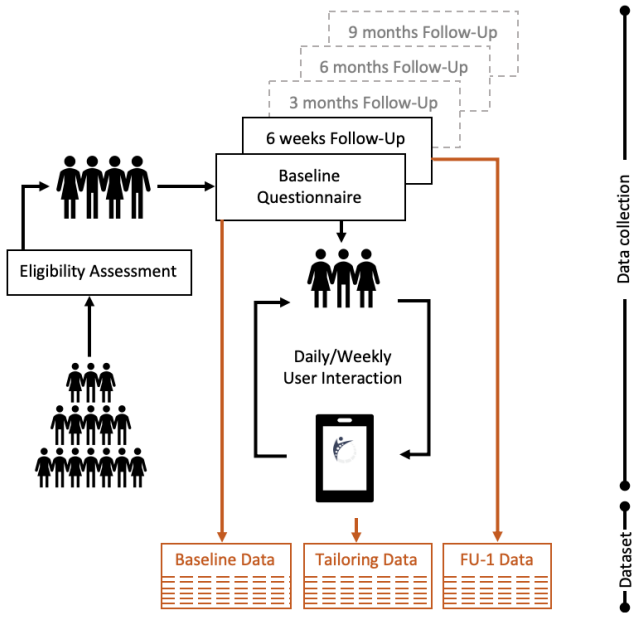


Figure 1: Overview of data collection in the selfBACK randomized controlled trial. The different data components are indicated by the orange boxes.

C.I Dataset 1

Dataset 1 consists of PROMs collected from LBP patients recruited in the intervention group of the selfBACK randomised controlled trial (RCT) [1], which aimed at facilitating self-management among patient with non-specific LBP.

Figure 1 shows the data collection in selfBACK. The data is categorised into Baseline, Tailoring and Follow-Up (FU) data. Patients were recruited through the referral of their primary care clinician, followed by screening for eligibility based on a set criteria. Eligible patients who accepted to join the study answered questionnaires at different time points: (1) at the time of intake: Baseline questionnaire (*Baseline Data*), (2) at the end of every week: Tailoring questionnaire (*Tailoring Data*), (3) at the end of 6-weeks, 3-months, 6-months, 9-months: Follow-Up questionnaire ((*FU Data*)). The

[1] *https://clinicaltrials.gov/ct2/show/NCT*03798288

questionnaire measures are:

- Pain level
- Pain self-efficacy
- Physical activity
- Sleep quality
- Fear avoidance
- Functional ability
- Work-ability
- Mood

The baseline questionnaire also included demographics (education, employment and family). The tailoring and follow-up questions are subsets of the baseline questions. A comprehensive overview of the data collection can be found in Sandal et al. (2019).

Based on the patients' responses at baseline, the selfBACK mobile application recommends an exercise plan and educational elements along with tracking their number of steps everyday from a wearable device (Xiaomi Mi Band 3). Exercise completion and educational readings were self-reported in the app. From this dataset, we only use the Baseline and FU-1 data for the experiments.

Target Outcomes

The average pain (last week) and work-ability reported by the patients in the FU-1 dataset were chosen as target outcomes from *dataset 1*, referred to as PA_f and WAI_f respectively. Average pain is self-assessed using the Numerical Pain Rating Scale (Hartrick et al., 2003), ranging from 0(*no pain*) to 10(*(disabling) severe pain*). Pain rating scales are commonplace in the medical and healthcare context and are used widely in different medical environments as a tool of communicating or expressing level of pain experienced by an individual. Work Ability Index (WAI) is a self-assessment measure used in workplace and occupational health surveys and uses the Numerical Rating Scale ranging from 0(*completely unable to work*) to 10(*workability at its best*) (Tuomi et al., 2002). It is widely used

in occupational health and research to facilitate understanding different dimensions of a working individual including their current ability to work compared with their lifetime best, self-prognosis of their work-ability in the last two years, their ability to work with respect to the demands of the job, the number of sick leaves taken in the last year, among others.

The dataset for predicting PA_f contains completed data from 218 patients, while for predicting WAI_f contains data from 159 patients. The number of patients is less in WAI_f due to the exclusion of patients who did not answer the baseline WAI, among them are the retired patients as this measure does not apply to them. The final dataset comprises of 47 self-reported measures, which form the predictor variables.

C.II Dataset 2

Data was collected by the Roessingh Center of Rehabilitation (RCR), Netherlands, between 2012-2019. The data consists of PROMs collected from NLBP patients referred to MBR using questionnaires administered at four time points: 1) before intake, 2) at the start, 3) at the end, and 4) after 3 months of pain rehabilitation, see figure 2. Patients gave consent to use their data for scientific research.

The questionnaires contain self-reported measures commonly used in pain rehabilitation,

- Hospital Anxiety and Depression Scale (HADS) (Bjelland et al., 2002)
- Multidimensional Pain Inventory (MPI) (Verra et al., 2012)
- Pain Disability Index (PDI) (Soer et al., 2013)
- Psychological Inflexibility in Pain Scale (PIPS) (Trompetter et al., 2014)
- Rand-36 Health Survey (RAND-36) (Saimanen et al., 2019)

The responses on the 121 questions were used to calculate 23 scores, shown in table 1. These scores are used as features in the ML experiment.

Target Outcome
The targets were the referral advice, which were given after the eligibility assessment (figure 2). The data set contained 1040 patient records. These records were labelled according to 4 possible referral advises:

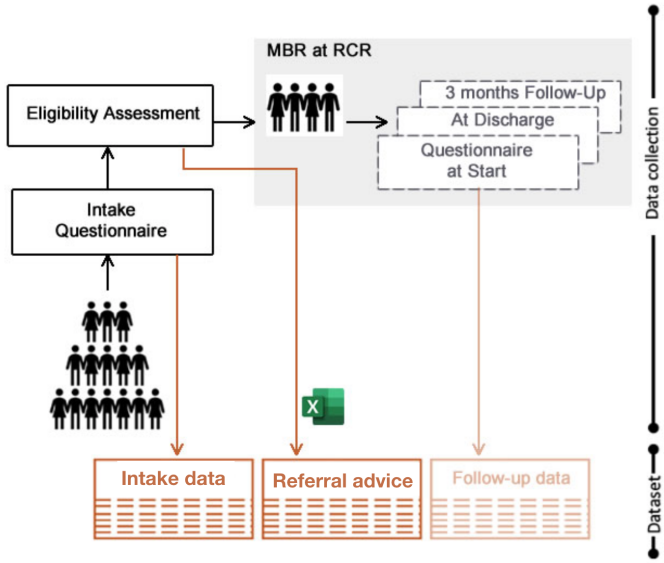


Figure 2: Overview of the assessment moments with questionnaires at the pain rehabilitation centre RCR, Enschede, the Netherlands. MBR: Multidisciplinary Biopsychosocial Rehabilitation.

Table 1: The PROMs included in Dataset 2

HADS		
Anxiety	Depression	Total score
MPI		
Pain severity	Interference	Life control
Affective distress	Solicitous responses	Distracting responses
Punishing responses	Support	Household chores
Outdoor work	Social activities	General activities
PDI		
Total score		
PIPS		
Avoidance	Cognitive fusion	Total score
RAND-36		
Physical functioning	Role limitations	Vitality
Mental health		

Table 2: Referral combinations the classification algorithms were trained on.

Model	Class A	Class B	# of cases
1	Clinic RCR	Polyclinic RCR	529
2	Clinic RCR	Reject	606
3	Polyclinic RCR	Reject	665
4	Polyclinic RMCR	Clinic RCR	375
5	Polyclinic RMCR	Polyclinic RCR	434
6	Polyclinic RMCR	Reject	511

– Clinic RCR (n=235): accepted for MBR at the RCR and advised to follow a clinical treatment path.

– Polyclinic RCR (n=294): accepted for MBR at the RCR and advised to follow a polyclinical treatment path.

– Polyclinic RMCR (n=140): referred to Roessingh Medinello Center of Rehabilitation (RMCR), which is similar to Polyclinic RCR but provides treatment paths for less complicated patients.

– Reject (n=371): referred to the RCR from primary or secondary care, but rejected after intake by clinician at RCR because they were not eligible.

This labelling resulted into an unbalanced dataset. The final dataset is shown in Table 2. The column # *of cases* shows the total number of cases (Class A + Class B) in Dataset 2 per combination.

D Methods

This section describes the ML tasks and the steps undertaken in the experiments. The ML pipeline used in this work is illustrated in figure 3.

D.I Regression

This task explores the application of different methods to determine which PROMs are optimal for predicting the target outcomes in *dataset 1* and

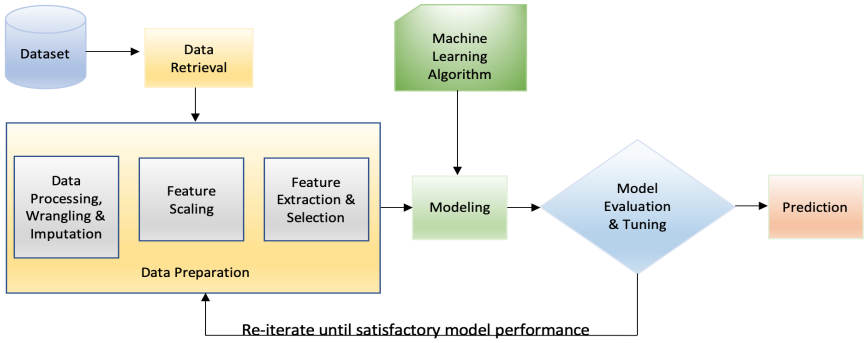


Figure 3: The workflow of the Machine Learning pipeline used in this study.

different supervised ML methods to determine the predictability of the outcomes and the best suited algorithm for this task.

Seven algorithms were used to estimate the target outcomes: *Linear Regression* (Driver and Kroeber, 1932), *Passive Aggressive Regression* (Crammer et al., 2006), *Random Forest Regression* (Svetnik et al., 2003), *Stochastic Gradient Descent Regression* (Robbins and Monro, 1951), *AdaBoost Regression* (Freund and Schapire, 1997), *Support Vector Regression* (Boser et al., 1992), *XGBoost Regression* (Chen and Guestrin, 2016). The algorithms were chosen based on the existing literature applying machine learning methods on PROM datasets in a bid to predict patient-specific outcomes (Rahman et al., 2018, 2019; Huber et al., 2019) and a number of experiments carried out previously where several algorithms were evaluated for their ability to predict patient-reported outcomes, including the ones mentioned above along with Neural Networks, k-NN, Gradient Boosting Machines among others, on similar regression tasks. The evaluation resulted in the selection of the above-mentioned seven algorithms, identified as most suitable for this task.

D.II Classification

We explored different ML methods to determine which PROMs are most useful for both referral of patients in- and to MBR using *dataset 2*. We used the clinician's decision as ground truth. Two classifier algorithms: 1) Balanced Random Forest (RF) classifier (Chen et al., 2004) and 2) Random Under-sampling Boosting classifier (RUSBoost) (Seiffert et al., 2009) were

chosen because of their ability to deal with class imbalance, handle small data sets and ease of interpretability. Both algorithms create an ensemble of models with a Decision Tree (Loh, 2011) as base estimator, which is also a classifier that has often been used in related work (Gross et al., 2013; Mamprin et al., 2020; D'Alisa et al., 2006). In addition, the respective classifiers were chosen because of their (1) integrated solution to deal with class imbalance; (2) ability to handle mixed data types; (3) ability to perform well with a small sample size (n≈1000); (4) high level of model interpretability; and (5) resemblance of thinking compared to a multidisciplinary team of health care professionals (Chen et al., 2004; Loh, 2011).

Binary classification tasks were created for the different referral combinations of the 1040 labelled samples, as shown in Table 2. Therefore, each classifier led to six models corresponding to the referral combinations. A nested cross validation was used to evaluate the performance of the models (Raschka, 2018). The nested cross-validation is a nesting of two k-fold cross-validation loops, with k representing the number of folds. The number of folds for both outer and inner loop was chosen to be 5, which is a very common number of folds for cross-validations. In other words, in every loop and for each binary classification task, data was divided into a training dataset with 80% of the samples, and a testing or validation dataset with 20% of the samples.

D.III Feature Selection

Feature selection becomes necessary for datasets with a large number of features to reduce the dimensionality without the loss of any important information. Reducing the dimensionality of the dataset before applying ML methods enables the algorithms to train faster by removing redundant information, thereby reducing the complexity and risk of overfitting the model(Chandrashekar and Sahin, 2014). Feature selection methods are broadly divided into three types: *filter*, *wrapper*, and *embedded*. *Filter methods* use the principal criteria of the ranking technique for selecting the most relevant features. Features are ranked based on statistical scores, such as correlation, to determine the features' correlation with the outcome variable. These methods are computationally efficient and do not rely on learning algorithms that can introduce a biased feature subset due to over-fitting (Chandrashekar and Sahin, 2014). However, a disadvantage of the filter method is that it does not consider the co-linearity among features in

the subset. Furthermore, it is difficult to precisely determine the dimension of the optimal feature subset (Chandrashekar and Sahin, 2014). *Wrapper methods* use the model's performance metric, for example accuracy, as an objective function to evaluate the feature subset (Chandrashekar and Sahin, 2014). These methods consider the association among features but are often too computationally expensive to perform an exhaustive search of the feature space. In *Embedded methods*, feature selection is integrated with the training progress of the model to reduce computational time compared to wrapper methods, while still considering the association among features (Guyon et al., 2008; Chandrashekar and Sahin, 2014). These methods iteratively extract features that contribute the most to the training for a particular iteration of a model during the training process. Regularisation methods (Santosa and Symes, 1986; Tibshirani, 1996) are commonly used embedded methods that penalise a feature based on a coefficient threshold. Feature Importance with ensemble methods is another method to determine impurity-based important features in tree-based algorithms [2]. Based on the trends observed in the existing literature, it was decided to use mutual information (only in classification task) (Ross, 2014) and impurity-based methods (Wittkowski, 1986; Fratello and Tagliaferri, 2018) in this work for selecting feature subsets.

D.IV Hyperparameter Optimization

Hyperparameter optimization is useful to find a set of hyperparameters that optimizes the performance of the algorithm (Claesen and De Moor, 2015). We considered model-based as well as model-free methods for hyperparameter optimization. Model-based optimization methods like Bayesian optimization use a surrogate model and an acquisition function to find the optimal set of hyperparameters (Yao et al., 2018; Hutter et al., 2019). We did not choose model-based optimization since the surrogate model is prone to overfitting on the hyperparameters (Lévesque, 2018) and this approach is more suitable to models that are computationally expensive to train, such as Deep Neural Networks (Hutter et al., 2019). Model-free methods can be categorized as heuristic and simple search approaches. Heuristic search approaches maintain a number of hyperparameter sets and use local perturbations and combinations of members in these sets to obtain an improved hyperparameter set (Yao et al., 2018; Hutter et al., 2019). Two common model-free simple

[2] *https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.htm*

search approaches are grid and random search (Yao et al., 2018). Grid search is one of the several ways of hyperparameter tuning and entails an exhaustive search through a defined set of hyper-parameter space of a learning algorithm. Random search selects the parameters at random instead of performing an exhaustive search over the hyperparameter space. We used random search in the classification task and grid search in the regression task to tune the hyperparameters of the algorithms.

D.V Evaluation Metrics

The evaluation metrics are different for each task owing to the very nature of different approaches undertaken. The evaluation metrics in the regression task are Mean Absolute Error (MAE), R-squared score (R^2) and Mean Residual (MR), while for the classification task are Matthews Correlation Coefficient (MCC) (Boughorbel et al., 2017), Balanced Accuracy (BAC) (Brodersen et al., 2010), Sensitivity (SEN) and Specificity (SPE) (Sammut and Webb, 2017). MAE is the average of the absolute errors, that is the difference between the observed value and the predicted value. R^2 is a goodness-of-fit metric to measure the proportion of variance explained by the independent variable(s) for a dependent variable in a regression model with values in the range [0,1], where 0 implies no observed variance and 1 implies 100% variance in the dependent variable with the movement of the independent variable(s). MR is the average difference between the predicted values and the observed values and is used to determine whether the models are likely to underestimate or overestimate the target value. MCC has a value in the range [-1,1] and produces a high score only when the predictions obtain good results in all of the four confusion matrix categories, which is useful for imbalanced classes (Chicco and Jurman, 2020). The value of BAC lies in the range of [0,1] and is a recommended metric for imbalanced classes (Brodersen et al., 2010). The values of SEN and SPE metrics lie in the range of [0,1] and are used widely to test the performance of binary classification models, where SEN is a measure of the proportion of correctly identified positives (true positive rate) while SPE is a measure of the proportion of correctly identified negatives (true negative rate).

E Experiments & Results

The experiments were done in Python (Oliphant, 2007) using Scikit-learn (Buitinck et al., 2013) and Imbalanced-learn (Lemaître et al., 2017) (only in classification task). k-fold cross validation is used in the experiments to reduce overfitting and increase the generalizability of the models, with $k = 5$ for classification and $k = 10$ for regression task.

E.I Regression Task

We used the embedded *feature importance* method of *Random Forest* algorithm to select the relevant features. Four and two features were selected for PA_f and WAI_f, respectively, which were then used to train the ML algorithms mentioned in the Methods section. The results are summarised in Table 3.

Table 3: Impurity-based feature selection using Random Forest for predicting PA_f (3a) and WAI_f (3b). The best performing model are highlighted in bold letters.

(a) PA_f

Model	**MAE±SD**	R^2	**MR**
LR	1.54±1.18	0.25	0.050
PAR	1.54±1.19	0.25	-0.087
SGDR	1.55±1.17	0.25	0.143
RFR	1.57±1.13	0.25	0.199
ABR	1.60±1.14	0.23	0.0
SVR	**1.53±1.15**	**0.27**	**0.102**
XGB	1.55±1.13	0.26	-0.015

(b) WAI_f

Model	**MAE±SD**	R^2	**MR**
LR	1.16±1.12	0.27	0.003
PAR	1.10±1.14	0.28	-0.288
SGDR	1.10±1.13	0.29	-0.243
RFR	**1.09±1.20**	**0.25**	**-0.246**
ABR	1.21±1.20	0.18	-0.090
SVR	1.11±1.15	0.27	-0.221
XGB	1.18±1.12	0.25	0.016

E.II Classification Task

We used the embedded feature selection method in both classifiers to select optimal features. For each classifier, six binary classification models were

Table 4: Results for the Balanced Random Forest (RF) classifier (± standard deviation).

	Train	**Test**			
	MCC	**MCC**	**BAC**	**SEN**	**SPE**
Model 1: C-P	0.22 ± 0.02	0.14 ± 0.08	0.56 ± 0.04	0.66 ± 0.13	0.47 ± 0.17
Model 2: C-R	0.26 ± 0.01	0.20 ± 0.08	0.60 ± 0.04	0.73 ± 0.11	0.47 ± 0.05
Model 3: P-R	0.22 ± 0.03	0.19 ± 0.06	0.60 ± 0.03	0.59 ± 0.06	0.61 ± 0.02
Model 4: M-C	0.54 ± 0.01	0.46 ± 0.05	0.73 ± 0.03	0.86 ± 0.13	0.59 ± 0.11
Model 5: M-P	0.42 ± 0.02	0.42 ± 0.05	0.70 ± 0.03	0.99 ± 0.03	0.42 ± 0.07
Model 6: M-R	0.53 ± 0.01	0.49 ± 0.06	0.77 ± 0.03	0.98 ± 0.04	0.57 ± 0.05

trained on different referral combinations, as shown in Table 2. The results are presented in Table 4 and Table 5.

The following observations were made based on the results:

- The overfit is low based on the MCC scores (both classifiers), except for the case Clinic RCR – Polyclinic RCR.
- The cases Polyclinic RMCR - Clinic RCR, Polyclinic RMCR - Polyclinic RCR and Polyclinic RMCR - Rejected show sub-optimal performances with their MCC's ranging between [0.42, 0.49] for RF and [0.43-0.50] for RUSBoost. Furthermore, their BAC scores are ranging between [0.70, 0.77] for RF and [0.71, 0.78] for RUSBoost.
- The cases Clinic - Rejected, Clinic RCR - Polyclinic RCR and Polyclinic RCR - Rejected all show very poor performances with their MCC's ranging between [0.14, 0.20] for RF and [0.11, 0.21] for RUSBoost. Furthermore, their BAC scores are ranging between [0.56, 0.60] for RF and [0.55, 0.60] for RUSBoost.

F Discussion

Our experiments on *dataset 1* indicate that ML methods and data science techniques can be used to identify relevant PROMs features and enhance

Table 5: Results for the Random Under Sampling Boosting (RUSBoost) classifier ± standard deviation.

	Train	**Test**			
	MCC	**MCC**	**BAC**	**SEN**	**SPE**
Model 1: C-P	0.22 ± 0.02	0.11 ± 0.07	0.55 ± 0.03	0.72 ± 0.10	0.39 ± 0.13
Model 2: C-R	0.24 ± 0.01	0.21 ± 0.08	0.60 ± 0.04	0.59 ± 0.16	0.61 ± 0.10
Model 3: P-R	0.20 ± 0.02	0.19 ± 0.06	0.60 ± 0.03	0.59 ± 0.06	0.61 ± 0.02
Model 4: M-C	0.55 ± 0.02	0.49 ± 0.10	0.74 ± 0.05	0.94 ± 0.13	0.54 ± 0.10
Model 5: M-P	0.43 ± 0.01	0.43 ± 0.05	0.71 ± 0.03	1.00 ± 0.00	0.42 ± 0.07
Model 6: M-R	0.52 ± 0.01	0.50 ± 0.05	0.78 ± 0.03	0.98 ± 0.03	0.57 ± 0.06

the prediction of patient outcomes, such as pain and work-ability. While in experiments using *dataset 2*, we found that the classifiers perform poorly in predicting treatment referral. These contrasting findings may be attributed to the different predictors available in the two datasets, their strength of association with the target outcomes or the fact that *dataset 1* had the target outcomes measured at baseline while *dataset 2* does not since it's a one time outcome given by the clinician.

F.I Clinical Relevance

To support shared clinical decision making, it is necessary to build prognostic models that can provide information to clinicians and patients of likely outcomes related to a certain treatment or symptoms profile.

In *dataset 1*, the baseline measurements of the associated target outcomes were their first most important predictors. The superior predictive value of baseline measurements of target outcomes has also been confirmed in other similar studies, such as by Fontana et al. (2019) and Huber et al. (2019). In *dataset 2*, the PROMs had low predictive power with regards to referral advises, which is similar to findings in our previous work (Nijeweme-d'Hollosy et al., 2018; Oude Nijeweme - d'Hollosy et al., 2020). Our results again emphasize the difficulty of referring NLBP patients based on PROMs and the need for more research on PROMs to include them in decision support on treatment referral.

F.II Data Science Relevance

From a data science perspective, PROM-based analytics is relatively uncharted territory, posing a unique challenge and presenting an opportunity for more research to test the existing methods and develop new ones that can facilitate furthering our comprehension of subjective datasets and their utility in improving patient-centred care. Building a comprehensive view of the patients using data-driven methods and evidence-based research can help clinicians and patients alike get practical insights from the available data to make shared strategic decisions. There is a need to increase awareness, availability, and understanding of subjective patient-centred data to build more sustainable and secure data ecosystems and facilitate a shift towards targeted interventions with the development of diagnostic and prognostic learning models.

G Conclusion and Future Work

The results presented in this work support our premise that the analytical abilities of ML methods can be leveraged for making predictions using PROMs, given that the PROMs hold predictive power. With better predictors, further development, and thorough validation, ML models can facilitate a shared decision-making process for patients with musculoskeletal disorders in clinical settings. Support Vector Machines, Random Forest, and Random Under-sampling Boosting methods delivered the best performance in the experiments and present promising potential for adaptability and utility in clinical practice. The biggest strength of ML methods is their ability to handle big data and their adaptability to different clinical setups where a certain level of accuracy is required to predict outcomes. There is, however, a need for the development of a standard ML pipeline to guide further research on developing as well as reporting results of ML models that can predict PROMs in other clinical or healthcare datasets with patient-reported outcomes.

Bibliography

Ingvar Bjelland, Alv A Dahl, Tone Tangen Haug, and Dag Neckelmann. The validity of the hospital anxiety and depression scale: an updated literature review. *Journal of psychosomatic research*, 52(2):69–77, 2002.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6), 2017.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE, 2010.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

Chao Chen, Andy Liaw, Leo Breiman, et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12):24, 2004.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

Marc Claesen and Bart De Moor. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.

Simonetta D'Alisa, G Miscio, S Baudo, A Simone, L Tesio, and Alessandro Mauro. Depression is the main determinant of quality of life in multiple sclerosis: a classification-regression (cart) study. *Disability and rehabilitation*, 28(5):307–314, 2006.

Harold E. Driver and A. L. Kroeber. *Quantitative expression of cultural relationships*. University of California Press, 1932.

Mark Alan Fontana, Stephen Lyman, Gourab K Sarker, Douglas E Padgett, and Catherine H MacLean. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research*, 477 (6):1267–1279, 2019.

Michele Fratello and Roberto Tagliaferri. Decision trees and random forests. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, page 374, 2018.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Douglas P Gross, Jing Zhang, Ivan Steenstra, Susan Barnsley, Calvin Haws, Tyler Amell, Greg McIntosh, Juliette Cooper, and Osmar Zaiane. Development of a computer-based clinical decision support tool for selecting

appropriate rehabilitation interventions for injured workers. *Journal of occupational rehabilitation*, 23(4):597–609, 2013.

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

Alex HS Harris, Alfred C Kuo, Yingjie Weng, Amber W Trickey, Thomas Bowe, and Nicholas J Giori. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research*, 477(2):452, 2019.

Craig T. Hartrick, Juliann P. Kovan, and Sharon Shapiro. The numeric rating scale for clinical pain measurement: A ratio measure? *Pain Practice*, 3(4): 310–316, 2003. doi: 10.1111/j.1530-7085.2003.03034.x.

Michelle M Holmes, George Lewith, David Newell, Jonathan Field, and Felicity L Bishop. The impact of patient-reported outcome measures in clinical practice for pain: a systematic review. *Quality of Life Research*, 26 (2):245–257, 2017.

Manuel Huber, Christoph Kurz, and Reiner Leidl. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC medical informatics and decision making*, 19(1):3, 2019.

Eric L Hurwitz, Kristi Randhawa, Hainan Yu, Pierre Côté, and Scott Haldeman. The global spine care initiative: a summary of the global burden of low back and neck pain studies. *European Spine Journal*, 27(6): 796–801, 2018.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning*. Springer, 2019.

Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in

machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

Julien-Charles Lévesque. Bayesian hyperparameter optimization: overfitting, ensembles and conditional spaces. 2018.

Ivan Lin, Louise Wiles, Rob Waller, Roger Goucke, Yusuf Nagree, Michael Gibberd, Leon Straker, Chris G Maher, and Peter PB O'Sullivan. What does best practice care for musculoskeletal pain look like? eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. *British journal of sports medicine*, 54(2):79–86, 2020.

Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

Gustavo C Machado, Marina B Pinheiro, Hopin Lee, Osman H Ahmed, Paul Hendrick, Chris Williams, and Steven J Kamper. Smartphone apps for the self-management of low back pain: A systematic review. *Best Practice & Research Clinical Rheumatology*, 30(6):1098–1109, 2016.

Marco Mamprin, Jo M Zelis, Pim AL Tonino, Svitlana Zinger, and Peter HN de With. Gradient boosting on decision trees for mortality prediction in transcatheter aortic valve implantation. *arXiv preprint arXiv:2001.02431*, 2020.

Paul Jarle Mork and Kerstin Bach. A decision support system to enhance self-management of low back pain: protocol for the selfback project. *JMIR research protocols*, 7(7):e167, 2018.

Wendy Oude Nijeweme-d'Hollosy, Lex van Velsen, Mannes Poel, Catharina GM Groothuis-Oudshoorn, Remko Soer, and Hermie Hermens. Evaluation of three machine learning models for self-referral decision support on low back pain in primary care. *International journal of medical informatics*, 110:31–41, 2018.

Travis E Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

Wendy Oude Nijeweme - d'Hollosy, Lex van Velsen, Mannes Poel, Catharina Groothuis-Oudshoorn, Remko Soer, Patrick Stegeman, and Hermie Hermens. Applying machine learning on patient-reported data to model

the selection of appropriate treatments for low back pain: A pilot study. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020)*, volume 5: HEALTHINF, pages 117–124. SCITEPRESS, February 2020. ISBN 978-989-758-398-8. doi: 10.5220/0008962101170124. URL http://www.healthinf.biostec.org/. 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020, BIOSTEC ; Conference date: 24-02-2020 Through 26-02-2020.

Quazi Abidur Rahman, Tahir Janmohamed, Meysam Pirbaglou, Hance Clarke, Paul Ritvo, Jane M Heffernan, and Joel Katz. Defining and predicting pain volatility in users of the manage my pain app: Analysis using data mining and machine learning methods. *Journal of medical Internet research*, 20(11):e12001, 2018.

Quazi Abidur Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, and Joel Katz. Interpretability and class imbalance in prediction models for pain volatility in manage my pain app users: analysis using feature selection and majority voting methods. *JMIR medical informatics*, 7(4):e15601, 2019.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2), 2014.

Iina Saimanen, Viivi Kuosmanen, Dina Rahkola, Tuomas Selander, Jari Kärkkäinen, Jukka Harju, Samuli Aspinen, and Matti Eskelinen. Rand-36-item health survey: A comprehensive test for long-term outcome and health status following surgery. *Anticancer research*, 39(6):2927–2933, 2019.

Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017.

Louise Fleng Sandal, Mette Jensen Stochkendahl, Malene Jagd Svendsen, Karen Wood, Cecilie K Øverås, Anne Lovise Nordstoga, Morten Villumsen, Charlotte Diana Nørregaard Rasmussen, Barbara Nicholl, Kay Cooper, Per Kjaer, Frances S Mair, Gisela Sjøgaard, Tom Ivar Lund Nilsen, Jan Hartvigsen, Kerstin Bach, Paul Jarle Mork, and Karen Søgaard. An app-delivered self-management program for people with low back pain: Protocol for the selfback randomized controlled trial. *JMIR Res Protoc*, 8(12):e14720, Dec 2019. ISSN 1929-0748. doi: 10.2196/14720.

Fadil Santosa and William W Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

Bruno Saragiotto, Matheus Almeida, TiÃª Yamato, and Chris Maher. Multidisciplinary biopsychosocial rehabilitation for nonspecific chronic low back pain. *Physical Therapy*, 96, 12 2015. doi: 10.2522/ptj.20150359.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

Remko Soer, Albère JA Köke, Patrick CAJ Vroomen, Patrick Stegeman, Rob JEM Smeets, Maarten H Coppes, and Michiel F Reneman. Extensive validation of the pain disability index in 3 groups of patients with musculoskeletal pain. *Spine*, 38(9):E562–E568, 2013.

Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

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

Hester R Trompetter, Ernst T Bohlmeijer, Bianca Van Baalen, Marco Kleen, Albère Köke, Michiel Reneman, and Karlein MG Schreurs. The psychological inflexibility in pain scale (pips). *European journal of psychological assessment*, 2014.

K Tuomi, J Ilmarinen, A Jahkola, L Katajarinne, and A Tulkki. Work ability index. *Finnish Institute of Occupational Health*, 2002.

Martin L Verra, Felix Angst, J Bart Staal, Roberto Brioschi, Susanne Lehmann, André Aeschlimann, and Rob A de Bie. Reliability of the multidimensional pain inventory and stability of the mpi classification system in chronic back pain. *BMC musculoskeletal disorders*, 13(1):155, 2012.

K Wittkowski. Classification and regression trees-l. breiman, jh friedman, ra olshen and cj stone. *Metrika*, 33:128–128, 1986.

Quanming Yao, Mengshuo Wang, Yuqiang Chen, Wenyuan Dai, Hu Yi-Qi, Li Yu-Feng, Tu Wei-Wei, Yang Qiang, and Yu Yang. Taking human out of learning applications: A survey on automated machine learning. *arXiv preprint arXiv:1810.13306*, 2018.

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

Using Automated Feature Selection for Building Case-Based Reasoning Systems: An Example from Patient-Reported Outcome Measurements

Deepika Verma, Kerstin Bach, Paul Jarle Mork

Abstract

Feature selection for case representation is an essential phase of Case-Based Reasoning (CBR) system development. To (semi-)automate the feature selection process can ease the knowledge engineering process. This paper explores the feature importance provided for XGBoost models as basis for creating CBR systems. We use Patient-Reported Outcome Measurements (PROMs) on low back pain from the SELFBACK project in our experiments. PROMs are a valuable source of information that capture physical, emotional as well as social aspects of well-being from the perspective of the patients. Leveraging the analytical capabilities of machine learning methods and data science techniques for exploiting PROMs have the potential of improving decision making. This paper presents a two-fold approach employed on our dataset for feature selection that combines statistical strength with data-driven knowledge modelling in CBR and compares it with permutation feature selection using XGBoost regressor. Furthermore, we compare the performance of the CBR models, built with the selected features, with two machine learning algorithms for predicting different PROMs.

A Introduction

Patient-reported outcome measurements (PROMs)[1] are collected routinely in clinical settings and are designed to capture the patients' perception of their own health through structured questionnaires. By utilising machine learning methods and data science techniques, there is a large potential for PROMs to inform and improve clinical decision making (Wu et al., 2013). In the current work, we use PROMs on low back pain (LBP) as an example. Among patients seen in primary care, a specific cause of LBP can rarely be identified and the symptoms are most often diagnosed as being "nonspecific". This also highlights the multi-factorial nature of LBP, i.e., both genetic, physiological, social and psychological factors are likely to contribute to LBP. While an early and thorough assessment of LBP is recommended (for example, to detect cases at high risk of poor outcome) (Lin et al., 2020), there are currently no clinical decision support systems (CDSS) in use in clinical practice that can assist or improve such detection or predict the likely outcome for a patient.

[1] *https://www.hss.edu/proms.asp*

Case-Based Reasoning (CBR) systems are well suited for the task of CDSS (Bichindaritz and Marling, 2010) since the PROMs of the patients can be described in a case-base, a knowledge repository that can aid decision making (Andritsos et al., 2014). However, clinical datasets with PROMs usually contain several clinical measures, all of which may not necessarily be required for decision making and it is therefore necessary to be able to select optimal subset of features that can be used for building CBR systems to predict the patient outcomes and facilitate decision making (Floyd et al., 2008).

Retrieval of similar cases is an important phase in CBR systems, which relies on the case representation and similarity measures. Hence, the selection of the most relevant and important features can easen and simplify the development of the entire CBR system. The focus of this paper is the feature selection phase for building CBR systems from PROMs to predict patient outcomes. While the overall method can be applied to other domains, we will present our evaluation using a dataset with PROMs (described in section C) in this work.

We employ a two-fold approach on our dataset for feature selection that combines statistical strength with data-driven knowledge modelling in CBR and compare it with permutation feature selection using XGBoost regressor. Additionally, we compare the performance of the CBR models, built with the selected features, with two machine learning algorithms for predicting different PROMs.

B Related Work

PROMs are a valuable source of information and present opportunities for highly sophisticated analysis, but has only been exploited by a few studies in the context of leveraging analytical capabilities of machine learning methods. Rahman et al. (2018) used a total of 130 PROMs collected via their pain self-management mobile application ("Manage My Pain"). Using Random Forest, they showed that pain volatility levels at 6 months follow-up could be predicted with a 70% accuracy. In their followup work, Rahman et al. (2019) showed that similar level of accuracy (68%) could be obtained with just 9 features. In another study, Harris et al. (2019) used preoperative PROMs to predict whether or not a patient achieves a clinically important improvement in several pain- and function-related outcomes at 1-year post

knee arthroplasty. Using several supervised machine learning algorithms, they showed that similar performance can be achieved across different algorithms for the outcomes by varying the number of inputs.

Using the CBR methodology for clinical datasets has already proven useful in decision making (Holt et al., 2005). For building robust decision support CBR systems, sufficient description of the problem is necessary. Knowledge about the importance of various features in the dataset plays an important role in problem description for building CBR systems (Aamodt and Plaza, 1994). Xiong and Funk (2006) proposed an approach wherein they assessed the feature subset selection based on the performance of CBR models. Later on, the authors proposed a hierarchical approach to select feature subsets for similarity models (Xiong and Funk, 2010). They used individual cases to optimise the possibility distributions in the case base and features were selected based on the magnitude of their parameters in the similarity models. Similar to the feature-selection approach proposed by Li et al. (2009), we identify optimal feature subsets for our CBR system by iteratively building CBR systems with different feature subsets and evaluating the performance based on the predictions. While Li et al. used mutual information as a preset criterion for selecting feature subsets and evaluating the subsequent CBR systems, we used correlation. In their previous work, Li et al. (2006) combined feature reduction using rough set with case selection for handling large datasets. Similarly, Zhu et al. (2015) selected reduced feature sets through neighborhood rough set algorithm, a method that has been used widely for feature and case selection in CBR (Salamó and Golobardes, 2001; Salamo and Lopez-Sanchez, 2011).

C SELFBACK Dataset

The dataset consists of PROMs collected during the randomised controlled trial (RCT)[2] that tested the effectiveness of the SELFBACK [3] DSS (Sandal et al., 2019).

Care-seeking patients in primary care with non-specific LBP were recruited to the study. Patients were screened for eligibility based on a set of criteria. The eligible patients were invited to participate in the RCT and those who accepted the invite answered a baseline questionnaire. The participating

[2] *https://clinicaltrials.gov/ct2/show/NCT*03798288

[3] *http://www.selfback.eu*

patients were randomized into either intervention group or control group. The intervention group had access to the SELFBACK DSS mobile application and received tailored self-management plans weekly whereas the control group did not. The participants answered questionnaires at different time-points: (1) (only intervention group) at the end of every week: Tailoring questionnaire, and (2) at the end of 6-weeks, 3-months, 6-months and 9-months: Follow-up questionnaire. The questionnaires consist of measures of *pain intensity, pain self-efficacy, physical activity, functional ability, work-ability, sleep quality, fear avoidance* and *mood*. Additionally, the baseline questionnaire included patient sociodemographics (education, employment and family). Table 1 summarises the information collected from the participants at various time-points. We use the Baseline, Follow-up 1 (after 6 weeks) and Follow-up 2 (after 3-months) PROMs in our evaluation. A detailed account of data collection for the RCT can be found in Sandal et al. (2019).

Table 1: The SELFBACK dataset created consists of participant characteristics collected at different time points and includes a selection of PROMs.

Descriptive variables		
Patient Characteristics	Sociodemographics	
Primary Outcome Measure		
Roland Morris Disability Questionnaire		
Secondary Outcome Measures		
Pain Self-Efficacy Questionnaire	Fear Avoidance Belief Questionnaire	Pain Intensity
Brief Illness Perception Questionnaire	Saltin-Grimby Physical Activity Level Scale	
Global Perceived Effect		
Other Outcome Measures		
Workability	Health-related Quality of Life	Activity Limitation
Patient Health Questionnaire	Perceived Stress Scale	Sleep
Patient Specific Functional Scale	Paint Duration and frequency	Physical Activity
Exercise		

From the dataset, six outcomes were selected as target outcomes: Roland Morris Disability Questionnaire (RMDQ, range: [0,24]), Numeric Pain Rating Scale (NPRS, range: [0,10]), Work-ability index (WAI, range: [0,10]), Pain Self Efficacy Questionnaire (PSEQ, range: [0,60]), Fear Avoidance Belief Questionnaire (FABQ, range: [0,30]) and Global Perceived Effect Scale (GPE, range: [-5,+5]). The primary outcome, RMDQ, is used to evaluate the effect of the self-management app in the RCT. The other outcomes were chosen to elucidate the variation in LBP symptoms amongst the participants.

The intervention group dataset consists of PROMs from 218 participants while the control group dataset contains PROMs of 158 participants. Each

participant is initially described by 47 features. Only the participants who completed at least the first two follow-up questionnaires were included in this work.

D Feature Engineering for CBR systems

Feature selection is an important step in the process of developing CBR systems. Reducing the dimensionality of the data enables the algorithm(s) to train faster by removing redundant information, thereby reducing model complexity, risk of overfitting, better generalisation and aiding interpretability of the models (Chandrashekar and Sahin, 2014). This is especially pertinent for building CBR systems for datasets with a high dimensionality, such as healthcare-oriented datasets, to ensure focus on the relevant attributes and enhance explainability of the models. Nonetheless, the methodology we present can be used for other domains for feature selection since the principle here is determining the best representation of a dataset in order to learn a solution to a given problem. While we use a healthcare domain dataset, the methodology itself has a broader application.

We use both filter and embedded methods in this work to determine reduced sets of predictors for the target outcomes. *Filter methods* use the principal criteria of ranking technique to select the most relevant features. Features are ranked based on statistical scores, correlation in our case, to determine the features' correlation with the outcome variable. This method is computationally efficient and does not rely on learning algorithms which can introduce a biased feature subset due to over-fitting (Chandrashekar and Sahin, 2014). However, correlation-based feature selection has shortcomings if there is a high degree of mutual correlation in the feature set. *Embedded methods* on the other hand are algorithm-specific, iteratively extracting features which contribute the most to the training of a particular iteration of a model during the training process. Impurity-based feature selection using tree-based algorithms[4] is a commonly used embedded method. Permutation feature importance determines the influence of random permutation of each predictor's values on the model performance while still preserving the distribution of the feature (Fisher et al., 2019).

We experimented with two methodologies for selecting optimal predictors for each target outcome:

[4] $https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html$

1. **Correlation and CBR**: Using a two-step hybrid method that combines statistical strength with data-driven case modelling, we attempted to derive optimal predictors of the target outcomes by computing correlation and iteratively building CBR models using features derived from correlation. Here, similarity measure development and building case representation are important factors in evaluating the performance of the CBR models for each set of features.

2. **Permutation feature importance using XGBoost**: Features are selected by computing permutation feature importance (PFI) with XGBoost (XGB) algorithm based on an evaluation metric.

Both methodologies aim to select optimal feature sets based on the trade-off between model performance and model simplicity, that is, fewer features.

D.I Feature Selection and CBR System Optimization

To determine the optimal set of predictors for developing CBR systems, we experimented with two methodologies for selecting features: correlation-based and based on the feature importance of a XGBoost model. The features selected by both methodologies were used to build CBR systems for all the outcomes at both follow-up time-points. Additionally, we implemented Support Vector and XGB Regression models to compare and contrast the performance of the CBR systems. Figure 1 illustrates the process of feature selection methods we used.

The modeling of the CBR systems was done with the myCBR workbench (Bach and Althoff, 2012). The experiments were run using myCBR Rest API[5] (Bach et al., 2019) for batch querying the CBR systems and python packages such as Scikit learn (Buitinck et al., 2013) and XGBoost (Chen and Guestrin, 2016) (python version 3.6.7) were used for building regression models and Pingouin for the statistical correlation (Vallat, 2018). For each target outcome we created datasets with the baseline data as input features and the PROMs of follow-up 1 and follow-up 2 as target values. These datasets were used to build CBR systems in a data-driven manner and as training data in the other two regression algorithms. In all the CBR models built for various target outcomes in this work, local similarity modelling of the attributes has

[5] $https://github.com/ntnu-ai-lab/mycbr-sdk$

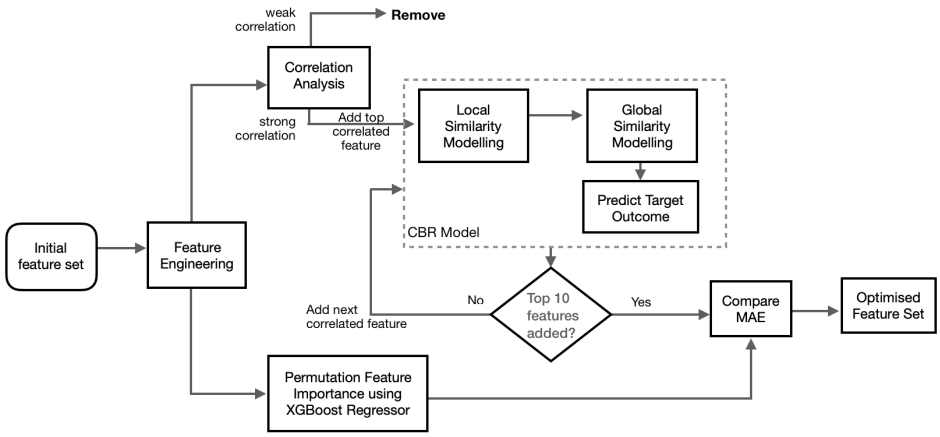
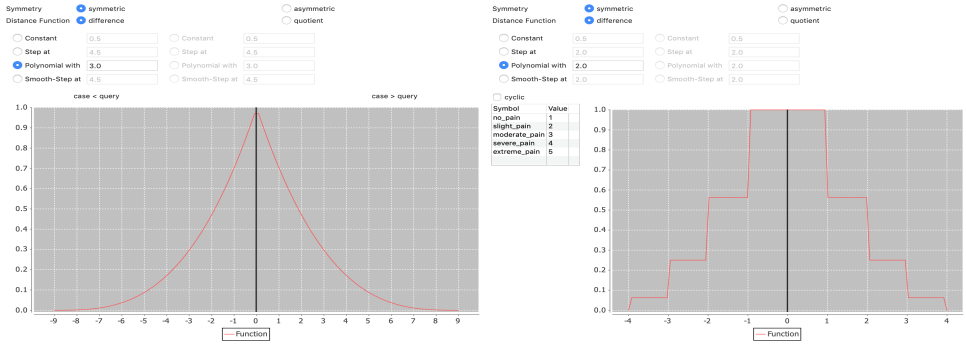


Figure 1: Flowchart of the feature selection process

been done in the same data-driven manner as presented in our previous work (Verma et al., 2018, 2019). The individual features are weighted equally in the global similarity function. Figure 2 showcases examples of local similarity measure modelling for numerical and categorical (ordinal) attributes (using correlated features of NPRS at follow-up 2 as an example). We urge the reader to refer to the previous work to fully grasp how the local similarity measures have been developed, as it is not possible to include the details in this work. Figure 3 shows the case representation of the same target outcome (NPRS) in myCBR workbench with 10 most correlated features.

To predict the target outcomes for a given participant using CBR model, we exploit the *"similar problems have similar solutions"* principle of CBR. While the query participant has been left out (leave-one-out cross validation), we determine their *n-nearest neighbours* (most similar case) with n in range [1,20] and compute mean of the target value reported by the n-neighbours, which is assigned as prediction for the given participant. The process is repeated for each participant and each target outcome dataset at both follow-up time-points for both the intervention and control group. The mean absolute error (MAE) is used as the metric to evaluate the predictive performance of the models.



(a) Baseline NPRS (b) EQ5D Pain

Figure 2: Modelling of Local similarity measures for numerical (a) and categorical (b) attributes in myCBR workbench.

Instance information

Name	Patient1	

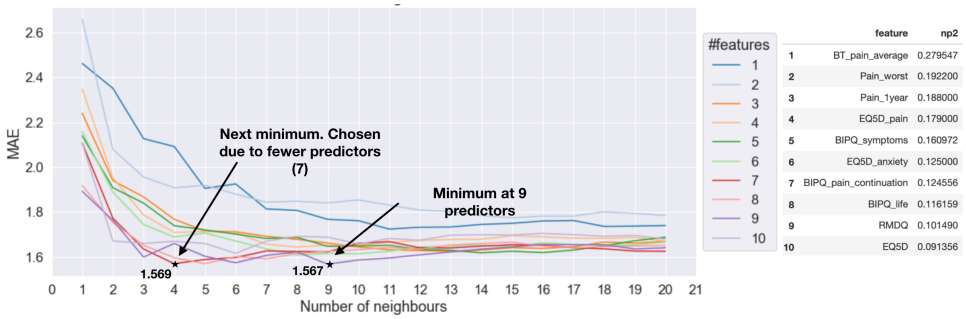
Attributes

BIPQ_life	6.0	Special Value: none
BIPQ_pain_continuation	10.0	Special Value: none
BIPQ_symptoms	5.0	Special Value: none
BT_pain_average	3.0	Special Value: none
EQ5D	80.0	Special Value: none
EQ5D_anxiety	not_anxious	Change Special Value: none
EQ5D_pain	slight_pain	Change Special Value: none
Pain_1year	Above30days	Change Special Value: none
Pain_worst	7.0	Special Value: none
RMDQ	7.0	Special Value: none

Figure 3: Case representation in myCBR for NPRS (at follow-up 2, control group dataset) with 10 most correlated features

D.II Correlation-based Feature Selection

Figure 1 shows that we we first compute correlation between the baseline features and each target outcome to select features. Since the dataset comprises of both numerical and categorical features, we use Pearson for numerical features and one-way ANOVA for categorical features to determine correlation between the baseline features and the target outcomes. Features with absolute correlation greater than the average correlation of the feature set and $p < 0.05$ were selected. For several reasons including simplified process of modelling in myCBR and based on experience from earlier



	feature	np2
1	BT_pain_average	0.279547
2	Pain_worst	0.192200
3	Pain_1year	0.188000
4	EQ5D_pain	0.179000
5	BIPQ_symptoms	0.160972
6	EQ5D_anxiety	0.125000
7	BIPQ_pain_continuation	0.124556
8	BIPQ_life	0.116159
9	RMDQ	0.101490
10	EQ5D	0.091356

Figure 4: On the right side of the figure are the top ten correlated features used to build the CBR model for predicting *NPRS* (input: baseline data, target: NPRS at follow-up 2). Features were added progressively one at a time in the given order, starting with the most correlated feature. **np2** (eta-squared) is the squared correlation coefficient. Graph on the left shows the MAE variation with different sets of features in the corresponding CBR model for predicting *NPRS*, with x-axis presenting the n-neighbours used for generating predictions and y-axis presenting the MAE in the predictions for the entire dataset.

experiments, it was decided to include only the top ten correlated features for building CBR systems. Previous experiments on the intervention group datasets showed that no more than ten features are necessary to predict any of the chosen target outcomes without any loss in the model performance. To build each CBR model, the casebase is populated with cases imported from a csv file in the myCBR workbench. Local similarity measures are developed for each attribute individually. Instead of building a new CBR model for each set of features, we build one model with the ten most correlated features and use ten different global similarity functions to progressively add more features. Once both the local and the global similarity measures are in place, we batch query the casebase using POST calls in the python implementation to generate predictions for the target outcome. The MAE is calculated between the reported outcome and the predictions for the entire dataset.

Figure 4 gives an example for one target outcome, *NPRS*. It shows the result of the correlation (left) and the MAE when predicting the NPRS using the baseline data (right). We can see that the progressive addition of correlated features improves the prediction by the CBR system already

by using the most similar case. Further, we observe that adding neighbors generally reduces the error and for the final model we choose the combination with the lowest MAE.

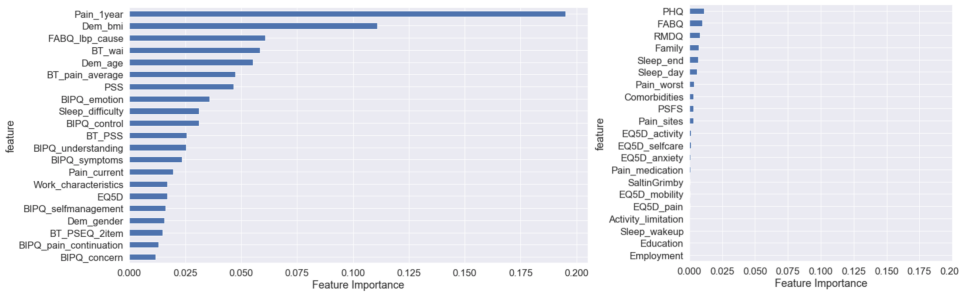
D.III Feature Importance using XGBoost

In this approach, we select features by computing the permutation feature importance using the XGBoost Regressor and compare the MAE of the predictions to determine the optimal feature set. The permutation feature importance is determined by the difference between the modified (permuted) dataset and a baseline model based on the MAE. First, a baseline model with all the features is trained and its MAE is computed. Next, the values of one feature in the dataset are permuted and then the model is re-trained and the MAE is computed for the modified dataset. The process is repeated for all the features in the dataset. The optimal number of features are selected based on the trade-off between model performance and number of features.

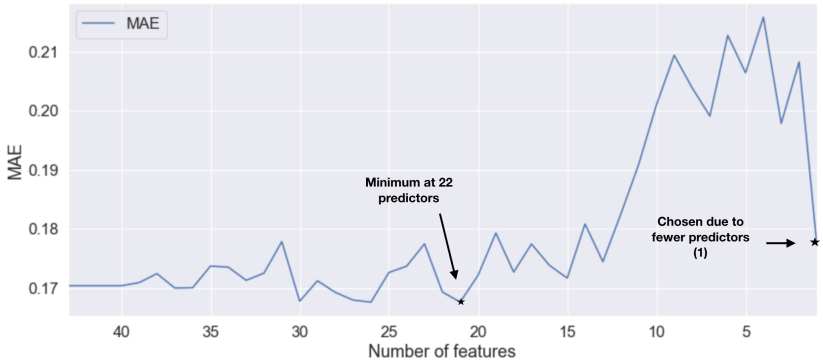
Figure 5a shows the feature importance for predicting the GPE and figure 5b shows the development of the MAE while adding the features. To select the best configuration, we choose the set with the lowest number of features that has the lowest MAE as shown in figure 5b. We favor the lowest number of features to build simpler model that requires minimal data collection and can be better explained. The selected features are then used to build CBR model in exactly the same way as described in the previous section and the prediction results are noted.

E Experimental Results

To compare the performance of the CBR systems, we implemented two regression algorithms, XGB and Support Vector Regression (SVR) for each corresponding CBR system to predict the target outcomes. The algorithms were selected based on previous experiments with the intervention group data where we evaluated the performance of XGB and SVR along with other algorithms, including Linear Regression, Passive Aggressive Regression, Stochastic Gradient Descent, AdaBoost, Random Forest, and found SVR and XGB to lead to the best results. For the simplicity of comparison and clarity, it was decided to keep only SVR and XGB for further evaluation. To optimize the hyperparameters, we used grid search (Hutter et al., 2019). Tables 2



(a)



(b)

Figure 5: Feature Selection using permutation feature importance with XGB for predicting *GPE* (input: baseline data, target: GPE at follow-up 1). **a.** Features ranked by their importance. **b.** Effect of feature permutation on the XGB model: The MAE on the y-axis in this plot is scaled.

and 3 summarise the results of predicting target outcomes using the CBR models, SVR and XGB for the intervention and control group participants, respectively.

F Discussion

A number of inferences can be made based on the results. We see in figure 4 that the baseline measurement (listed as *BT_pain_average*) of the associated target outcome *NPRS* is its' first most important predictor. This is a trend observed for all the target outcomes, except *GPE* which does not have an

Table 2: Results of Prediction of Target Outcomes using different Feature Selection Methodologies and Regression Methods for the Intervention Group (size of dataset: 218 participants). Numbers in bold letters are lowest MAE. **FU1**: Follow-up 1, **FU2**: Follow-up 2, **n**: number of features

		Feature Selection Methodology							
		Correlation+CBR				**PFI+XGBoost**			
Target	**Follow-Up**	**n**	**CBR**	**SVR**	**XGB**	**n**	**CBR**	**SVR**	**XGB**
RMDQ	**FU1**	4	2.98	3.19	3.32	5	2.78	**2.69**	2.71
	FU2	8	2.90	**2.83**	2.85	4	3.17	3.92	3.02
NPRS	**FU1**	7	**1.38**	1.45	1.50	3	1.50	1.49	1.52
	FU2	9	1.48	**1.33**	1.38	3	1.46	1.41	1.42
WAI	**FU1**	5	1.16	1.98	1.98	2	**1.14**	1.96	2.01
	FU2	4	**1.14**	2.16	2.21	1	1.24	2.19	2.24
PSEQ	**FU1**	1	5.50	16.9	17.0	2	**5.45**	17.2	17.3
	FU2	3	5.95	16.6	16.6	2	**5.95**	16.4	17.1
FABQ	**FU1**	3	3.87	3.74	3.76	6	3.90	**3.50**	3.67
	FU2	1	3.9	**3.60**	3.84	6	3.83	3.64	3.86
GPE	**FU1**	1	**1.37**	2.73	2.76	2	1.39	2.82	2.78
	FU2	2	1.54	2.51	2.43	3	**1.49**	2.54	2.46

associated baseline measurement (see figure 5b). This is an important observation from clinical perspective, since baseline measurements of the associated outcomes have previously been found to be their most important predictor (Fontana et al., 2019; Huber et al., 2019), and our experiments support these findings.

Selecting optimal features, especially for healthcare datasets, is one of those application domains where no one particular method prevails and one must decide based on application domain knowledge and experience, among others. From the results in table 2 and 3, we see that the features selected by either of the methodologies give similar results with respect to the error in predictions. There is no clear winner here. However, taking into consideration the time and effort required, XGBoost permutation feature importance methodology requires minima and provides a more streamlined process for

Table 3: Results of Prediction of Target Outcomes using different Feature Selection Methodologies and Regression Methods for the Control Group (size of dataset: 158 participants). Numbers in bold letters are lowest MAE. **FU1**: Follow-up 1, **FU2**: Follow-up 2, **n**: number of features

		Feature Selection Methodology							
		Correlation+CBR				**PFI+XGBoost**			
Target	**Follow-Up**	**n**	**CBR**	**SVR**	**XGB**	**n**	**CBR**	**SVR**	**XGB**
RMDQ	**FU1**	2	3.11	2.99	2.97	4	3.07	2.92	**2.75**
	FU2	2	3.11	**2.97**	3.14	3	3.22	2.97	3.14
NPRS	**FU1**	6	**1.41**	1.77	1.85	2	1.49	1.73	1.85
	FU2	7	1.56	**1.49**	1.7	1	1.72	1.56	1.71
WAI	**FU1**	1	1.02	1.02	**1.01**	1	1.02	1.02	1.01
	FU2	2	1.14	**1.12**	1.17	1	1.19	1.15	1.18
PSEQ	**FU1**	7	**6.68**	19.2	19.6	1	7.01	19.4	19.8
	FU2	3	6.23	19.0	19.5	5	**5.94**	19.1	19.3
FABQ	**FU1**	1	3.47	**3.27**	3.58	1	3.47	3.27	3.58
	FU2	2	3.77	3.69	3.80	2	3.82	**3.58**	3.93
GPE	**FU1**	7	**1.22**	2.55	2.52	1	1.26	2.61	2.49
	FU2	1	**1.33**	2.65	2.58	2	1.39	2.67	2.56

selecting optimal feature sets as compared to the two-fold approach, which requires estimating correlation, building several similarity measures and CBR systems for the target outcomes and comparing the MAE for determining optimal feature sets. As for a concrete time comparison, it is not possible since the modelling of local and global similarity measures for building a CBR model requires manual input. On the other hand, this comparison also establishes the utility of the two-fold approach for building tailored CBR systems.

All the three regression methods give fairly similar results when it comes to predicting the outcomes. However, for an outcome with a relatively large range (*PSEQ*) or no baseline measurement of the target outcome (*GPE*), both SVR and XGB fall short in comparison to the results we get from the CBR models. This is similar to our findings in our previous work (Verma

et al., 2018) where we found CBR model built with our data-driven modelling approach to be more sensitive and robust to the data-distribution of individual features, thereby, furthering our premise that both data-driven similarity modelling and CBR are better suited for this task. Moreover, outcomes generated by CBR models are more explainable, which is a pre-requite for any CDSS where explainable systems are preferred over complex ones.

G Conclusion and Future Work

In this paper, we presented a two-fold approach for feature selection wherein we used the correlation coefficient as a pre-processing step to select ten most correlated features and build the CBR models with progressively more features for predicting PROMs. We examine the performance of the predictions generated using CBR systems to determine optimal feature subsets for the outcomes. Through evaluation and comparison with tree-based feature selection methods (permutation feature importance with XGBoost), it can be concluded that although the presented two-fold approach is feasible and gives results similar to the other approach undertaken, it is however more time and effort intensive and therefore, feature selection using XGBoost permutation feature importance appears to be a more promising option. Predictive performance of the CBR systems is at par with and many a times better than the traditional algorithms such as SVR and XGBoost.

From a clinical perspective, building prognostic models that can provide necessary information to clinicians and patients of possible outcome(s) pertaining to a specific treatment is a necessity to support informed clinical decision making. Access to individualized predictive analytics for different outcomes may be the next step in the management of pain and related symptoms for patients with LBP. The results we get from our dataset confirm the predictive value of baseline measurements of associated target outcomes, similar to other studies such as by Fontana et al. (2019) and Huber et al. (2019).

In future work, it may be worthwhile to compare performance of the CBR models built with features selected by an expert with the approach presented in this work.

Bibliography

Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications,*, 7(1):39–59, 1994.

Periklis Andritsos, Igor Jurisica, and Janice I Glasgow. Case-based reasoning for biomedical informatics and medicine. In *Springer Handbook of Bio-/Neuroinformatics*, pages 207–221. Springer, 2014.

Kerstin Bach and Klaus-Dieter Althoff. Developing case-based reasoning applications using mycbr 3. In *International Conference on Case-Based Reasoning*, pages 17–31. Springer, 2012.

Kerstin Bach, Bjørn Magnus Mathisen, and Amar Jaiswal. Demonstrating the mycbr rest api. In *ICCBR Workshops*, pages 144–155, 2019.

Isabelle Bichindaritz and Cindy Marling. *Case-Based Reasoning in the Health Sciences: Foundations and Research Directions*, volume 309, pages 127–157. Springer, 05 2010. doi: 10.1007/978-3-642-14464-6$_7$.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785.

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. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Michael W Floyd, Alan Davoust, and Babak Esfandiari. Considerations for real-time spatially-aware case-based reasoning: A case study in robotic soccer imitation. In *European Conference on Case-Based Reasoning*, pages 195–209. Springer, 2008.

Mark Alan Fontana, Stephen Lyman, Gourab K Sarker, Douglas E Padgett, and Catherine H MacLean. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research*, 477(6): 1267–1279, 2019.

Alex HS Harris, Alfred C Kuo, Yingjie Weng, Amber W Trickey, Thomas Bowe, and Nicholas J Giori. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research*, 477(2): 452, 2019.

Alec Holt, Isabelle Bichindaritz, Rainer Schmidt, and Petra Perner. Medical applications in case-based reasoning. *Knowledge Engineering Review*, 20(3): 289–292, 2005.

Manuel Huber, Christoph Kurz, and Reiner Leidl. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC medical informatics and decision making*, 19(1): 3, 2019.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning*. Springer, 2019.

Yan Li, Simon Chi-Keung Shiu, Sankar K Pal, and James Nga-Kwok Liu. A rough set-based case-based reasoner for text categorization. *International journal of approximate reasoning*, 41(2):229–255, 2006.

Yan-Fu Li, Min Xie, and TN Goh. A study of mutual information based feature selection for case based reasoning in software cost estimation. *Expert Systems with Applications*, 36(3):5921–5931, 2009.

Ivan Lin, Louise Wiles, Rob Waller, Roger Goucke, Yusuf Nagree, Michael Gibberd, Leon Straker, Chris G Maher, and Peter PB O'Sullivan. What does best practice care for musculoskeletal pain look like? eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. *British journal of sports medicine*, 54(2):79–86, 2020.

Quazi Abidur Rahman, Tahir Janmohamed, Meysam Pirbaglou, Hance Clarke, Paul Ritvo, Jane M Heffernan, and Joel Katz. Defining and predicting pain volatility in users of the manage my pain app: Analysis using data mining and machine learning methods. *Journal of medical Internet research*, 20(11): e12001, 2018.

Quazi Abidur Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, and Joel Katz. Interpretability and class imbalance in prediction models for pain volatility in manage my pain app users: analysis using feature selection and majority voting methods. *JMIR medical informatics*, 7 (4):e15601, 2019.

Maria Salamó and Elisabet Golobardes. Rough sets reduction techniques for case-based reasoning. In *International Conference on Case-Based Reasoning*, pages 467–482. Springer, 2001.

Maria Salamo and Maite Lopez-Sanchez. Rough set based approaches to feature selection for case-based reasoning classifiers. *Pattern Recognition Letters*, 32 (2):280–292, 2011.

Louise Fleng Sandal, Mette Jensen Stochkendahl, Malene Jagd Svendsen, Karen Wood, Cecilie K Øverås, Anne Lovise Nordstoga, Morten Villumsen, Charlotte Diana Nørregaard Rasmussen, Barbara Nicholl, Kay Cooper, Per Kjaer, Frances S Mair, Gisela Sjøgaard, Tom Ivar Lund Nilsen, Jan Hartvigsen, Kerstin Bach, Paul Jarle Mork, and Karen Søgaard. An app-delivered self-management program for people with low back pain: Protocol for the selfback randomized controlled trial. *JMIR Res Protoc*, 8(12):e14720, Dec 2019. ISSN 1929-0748. doi: 10.2196/14720.

Raphael Vallat. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026, 2018. doi: 10.21105/joss.01026. URL `https://doi.org/10.21105/joss.01026`.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Modelling similarity for comparing physical activity profiles - a data-driven approach. In Michael T. Cox, Peter Funk, and Shahina Begum, editors, *CBR Research and Development*, Cham, 2018. Springer. ISBN 978-3-030-01081-2.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Similarity measure development for case-based reasoning–a data-driven approach. In Kerstin Bach and Massimiliano Ruocco, editors, *Nordic Artificial Intelligence Research and Development*, pages 143–148, Cham, 2019. Springer International Publishing. ISBN 978-3-030-35664-4.

Albert Wu, Hadi Kharrazi, L. Boulware, and Claire Snyder. Measure once, cut twice -adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. *Journal of clinical epidemiology*, 66:S12–20, 08 2013. doi: 10.1016/j.jclinepi.2013.04.005.

Ning Xiong and Peter Funk. Construction of fuzzy knowledge bases incorporating feature selection. *Soft Computing*, 10(9):796–804, 2006.

Ning Xiong and Peter Funk. Combined feature selection and similarity modelling in case-based reasoning using hierarchical memetic algorithm. In *IEEE Congress on Evolutionary Computation*, pages 1–6. IEEE, 2010.

Guo-Niu Zhu, Jie Hu, Jin Qi, Jin Ma, and Ying-Hong Peng. An integrated feature selection and cluster analysis techniques for case-based reasoning. *Engineering Applications of Artificial Intelligence*, 39:14–22, 2015.

Application of Machine Learning on Patient-Reported Outcome Measurements for Predicting Outcomes: A Literature Review

Deepika Verma, Kerstin Bach, Paul Jarle Mork

Abstract

The field of patient-centred healthcare has during the recent years adopted machine learning and data science techniques to support clinical decision making and improve patient outcomes. We conducted a literature review with the aim of summarising the existing methodologies that apply machine learning methods on patient-reported outcome measures datasets for predicting clinical outcomes to support further research and development within the field. We identified 15 articles published within the last decade that employ machine learning methods at various stages of exploiting datasets consisting of patient-reported outcome measures for predicting clinical outcomes. Furthermore, we discuss the gaps and challenges that can potentially be addressed in the future studies.

A Introduction

There is growing interest and support for the utility and importance of patient-reported outcome measures (PROMs) in clinical care. PROMs are commonly defined as reports or questionnaires completed by patients to measure their view on their functional well-being and health status (Kingsley and Patel, 2017). Thus, PROMs may capture the patient's perspective on both social, physical, and mental wellbeing. Shifting the focus from disease-specific factors towards the patient's perspective may provide a useful basis for shared medical decision-making between a clinician and a patient (Bingham III et al., 2017; Barry and Edgman-Levitan, 2012). Recent evidence indicates that shared decision-making has a positive impact on quality of decision-making, satisfaction with treatment, and patient-provider experience (Coronado-Vázquez et al., 2020). Likewise, well-informed patients agreeing upon their course of treatment with their caregiver have better outcome, and satisfaction (Sepucha et al., 2018).

PROMs may play an important role in shared decision-making, however, there is currently an unused potential in both collecting and utilising PROMs in clinical practice. Notably, digital innovations can facilitate delivery, storage, processing, and access to PROMs using third-party or electronic health record (EHR)-based outcome measurement platforms. Intelligent methods can also support shared decision-making through digital decision aids and patient engagement platforms comprising high-quality educational material, and

patient-provider communication portals (Jayakumar et al., 2017; Sepucha et al., 2018). In this context, utilising machine learning and artificial intelligence provide a promising avenue for enhancing the usefulness of PROMs (Giga, 2017).

Several recent studies demonstrated the predictive prowess of machine learning models utilizing EHR datasets for scheduling of surgeries (ShahabiKargar et al., 2014; Kargar et al., 2013; Devi et al., 2012), risk stratification (Wong et al., 2017; Moonesinghe et al., 2013; Marufu et al., 2016) among others. Singal et al. (2013) in their work found the machine learning models to outperform conventional models in predicting the development of hepatocellular carcinoma among cirrhotic patients. The application of machine learning methods on PROMs datasets can allow exploration of associations in the data that are important for predicting different outcomes and thereby inform a shared decision-making process (Mansell et al., 2021). Currently, PROMs data is widely used in explanatory research, where researchers typically test hypotheses using a preconceived theoretical construct by applying statistical methods (for example, low back pain is associated to lower quality of life and depression (Krismer et al., 2007; Waljee et al., 2014). In contrast, PROMs in predictive research can be used to predict outcomes in the future by applying statistical or machine learning methods without any preconceived theoretical constructs (for example, predicting the risk of depression (Andrews et al., 2017)), and is therefore an important step towards patient-centred care with a shift in focus towards the patient's perspective (Wang and Gottumukkala, 2020).

While prediction models exist that utilize a combination of PROMs and objective clinical data or EHR data for individual predictions (Baumhauer, 2017), models that utilise solely PROMs data to make individual predictions are rare. Despite the broad area of application of machine learning and data science techniques in the biomedical field, the utilisation of these techniques in clinical practice remains low, especially concerning the utilization of PROMs. A few machine learning applications utilizing PROMs data in biomedical research have emerged during recent years; however, the potential for utilising PROMs data to improve clinical care appears under-explored, especially from the perspective of supporting shared decision-making.

The main of aim of this literature review was therefore to provide a summary of existing methodologies that apply machine learning methods on PROMs for predicting clinical outcomes and building prognostic models. In Section B, we introduce the process of article selection and present an

analysis of the selected articles in terms of their publication year, intervention domains, length of outcome prediction, data source, feature selection strategy and the machine learning methods used. Furthermore, we discuss the gaps and challenges in Section C that can be addressed in future work to utilise machine learning methods on PROMs datasets. The main contribution of this work is firstly, identification of scientific articles applying machine learning methods on PROMs data for predicting clinical outcomes and secondly, augmenting the utility of machine learning methods for healthcare datasets for building clinical decision support systems to better facilitate decision making for patient-centred care and precision medicine.

B Methods

B.I Review Design and Search Strategy

This literature review identifies scientific articles that focus on the application of machine learning methods in the process of predicting short or long-term clinical outcome(s) using PROMs data.

A structured literature search was performed in September 2020 using the following search string in the PubMed and Scopus database: (((self reported measures) OR patient reported measures)) AND ((artificial intelligence) OR machine learning) AND ((outcome prediction) OR outcome assessment). The results were filtered to include journal and conference articles written in English and published within the last decade (2010–2020).

B.II Article Selection

The following inclusion criteria were used to identify articles relevant for the current review:

- *Data*: The dataset consists of structured questionnaires administered to patients or participants either in-person or via web application before, during and/or after a treatment. Articles that involved objectively measured data or data gathered from online patient forums were excluded from this study.

- *Machine Learning*: Application of machine learning methods with the intent of data analysis or clustering of patients or assessment

of features with prognostic value for one or more target outcomes or building prognostic models for short or long term prediction of one or more outcome.

- Full text availability (including institutional access)
- Written in English

Articles not meeting the inclusion criteria following the abstract and full screening were excluded from this study.

B.III Search Outcome

Figure 1 presents a flowchart of the article selection process. Based on the structured literature search, a total of 319 records were identified: PubMed (n = 314) and Scopus (n = 5). Further, we screened the references of the articles that met the inclusion criteria along with relevant review articles and books to identify additional articles (n = 4). Finally, after duplicates were removed, we screened 322 articles. After screening of title/abstract and assessing the eligibility, a total of 15 articles were included in the qualitative synthesis.

B.IV Sources of Evidence

All the included articles were published in peer-reviewed journals. Eight out of the 15 articles were published in the years 2019 and 2020 (excluding October–December 2020); see Figure 2. Fourteen articles were published the second half of the decade 2016–2020, while only one article was published in the first half of the decade, in 2012.

B.V Intervention Domains and Length of Prediction

Articles stratified by the intervention domain (Figure 3), can be broadly categorized as post-surgical improvements or limitations, depression, pain management, hospital readmission, and oral health.

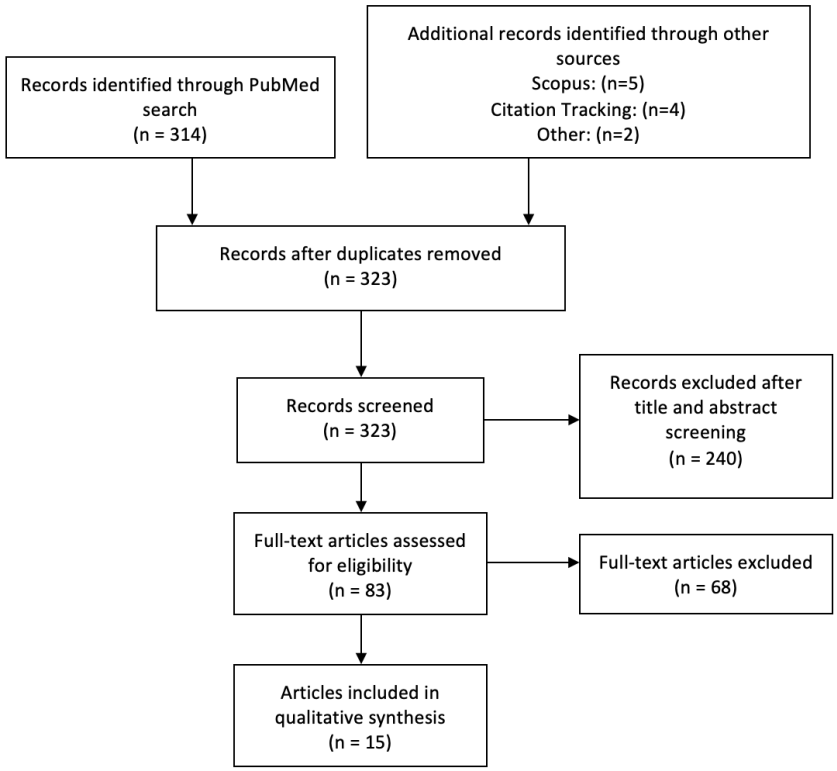


Figure 1: Flowchart of the article selection process.

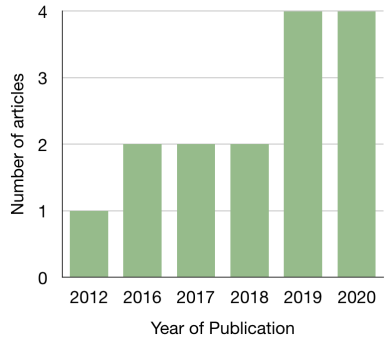


Figure 2: Publication year of included articles.

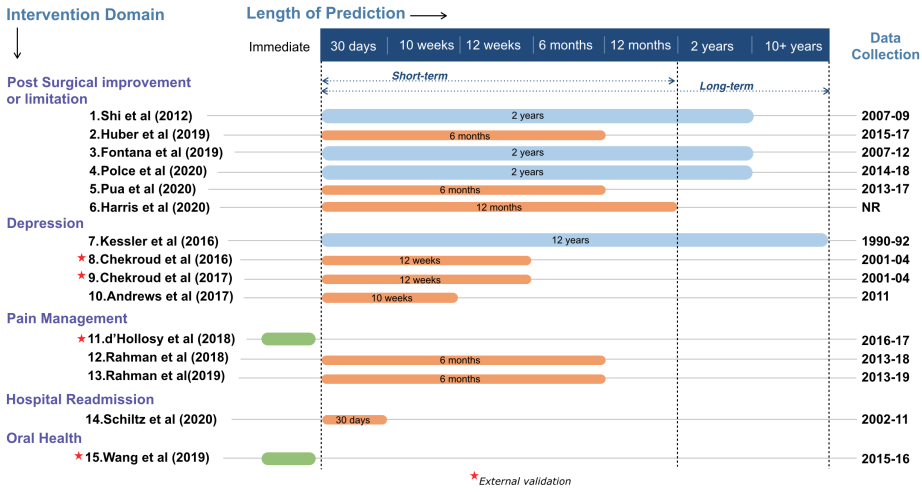


Figure 3: The included articles categorised by their intervention domains. The length of the predictions are indicated, categorised into short- and long-term. The time period of the data collection is indicated to the right. Red asterisks indicate studies that utilized external validation datasets to test the generalizability of the machine learning models.

The first category includes six articles, focusing on outcomes relating to post-surgical limitations or improvements such as quality of life after cancer surgery (Shi et al., 2012) and (walking) limitations or improvements (minimal clinically important difference (MCID)) after total joint arthroplasty (Huber et al., 2019; Pua et al., 2019; Fontana et al., 2019; Polce et al., 2020; Harris et al., 2019). The second category includes four articles, focusing on identifying patients with depression based on self-reports (Kessler et al., 2016; Andrews et al., 2017) and prognosis of outcome of anti-depression treatment (Chekroud et al., 2016, 2017). The third category includes three articles focusing on predicting pain volatility amongst users of a pain-management mobile application (Rahman et al., 2018, 2019) and self-referral decision support for patients with low back pain in primary care (Nijeweme-d'Hollosy et al., 2018). The fourth category includes one article that focused on the risk of hospital readmission (Schiltz et al., 2020), while the fifth and last category includes one article that focused on oral health outcome among children aged 2–17 years (Wang et al., 2020b).

Eleven articles presented machine learning models for predicting short-

term outcomes (12 months or less), see Figure 3, while four articles presented machine learning models for predicting long-term outcomes (over 12 months). Two articles focused on immediate outcomes, such as referral decision (Nijeweme-d'Hollosy et al., 2018) and oral health scores (Wang et al., 2020b). Four articles, marked with red asterisk in Figure 3, utilized external validation datasets to test the generalizability of the machine learning models. None of the articles with long-term outcomes utilised external validation datasets. The prediction timelines also appear to be domain dependant. Outcomes from interventions like depression treatment or surgeries seem to be predicted over long-term, likely due to the nature of the treatment and associated outcomes in the two intervention domains.

B.VI Sources of Data and Availability

Table 1 presents a summary of the included articles. Few articles utilized open-source or available-on-request datasets from national registries such as National Institute of Mental Health (NIMH) or National Health Service (NHS). The size of datasets vary, from 37 patients (Andrews et al., 2017) to 64,634 patients (Huber et al., 2019). Seven articles utilized training datasets with less than 1000 patients.

B.VII Feature Selection

The methods of feature selection were either statistical, algorithm-based or manual, based on expertise or availability of data (Table 1). In the table, 'Algorithm implicit' implies that the features were selected by the algorithm(s) used for the prediction task and no other explicit feature selection was carried out, while 'Manual' implies that the features were selected manually based on experience or expert knowledge or data availability.

Ten articles used supervised learning algorithms to extract relevant features from the dataset, while in four articles, features were selected manually, without any statistical or algorithmic assistance. One article (Shi et al., 2012) applied statistical methods to extract and select relevant features. Among the four articles that employed manual feature selection, two articles (Wang et al., 2020b; Fontana et al., 2019) manually divided all the features into sets and added the sets incrementally into the training dataset to train the model(s). In comparison, in the other two articles (Pua et al., 2019; Harris et al., 2019), features were selected manually based on clinical expertise

Table 1: Overview of feature selection, model evaluation, data availability and external validation in the included articles. Abbreviations- MCID: Minimal Clinically Important Difference, LBP:Low Back Pain, NR: Not Reported, ANOVA: Analysis of Variance, RFE: Recursive Feature Elimination, RF: Random Forest, CV: Cross Validation, RoC: Receiver operating characteristic, LASSO: Least Absolute Shrinkage and Selection Operator, NHS: National Health Service, NIMH: National Institute of Mental Health, HRS: Health and Retirement Study.

Article	Outcome	Dataset Size	Total No. of Features	Features Selected	Feature Selection Method	Hyperparameter Tuning	Model Evaluation	Data Availability	External Validation
Shi et al. (2012)	Quality of life post surgery	403	NR	NR	ANOVA, Fisher exact analysis, Univariate analysis	NR	Holdout (80,20)	NR	no
Huber et al. (2019)	MCID post-surgery	64,634	81	NR	Algorithm implicit	NR	5-fold CV	NHS [1]	no
Fontana et al. (2019)	MCID post-surgery	13,809	NR		Manual	5-fold CV	Holdout (80,20)	NR	no
Polce et al. (2020)	Satisfaction post-surgery	413	16	10	RFE, RF	10-fold CV	Holdout (80,20)	NR	no
Pua et al. (2019)	Walking limitation post-surgery	4026	NR	25	Manual	5-fold CV	Holdout (70,30)	NR	no
Harris et al. (2019)	MCID post-surgery	587	NR	NR	Manual	NR	10-fold CV, bootstrapping	NR	no
Kessler et al. (2016)	Depressive Disorder chronicity, persistence, severity	1056	NR	9–13	Ensemble,Penalized Regression	NR	10-fold CV	NR	no
Chekroud et al. (2016)	Antidepressant treatment	1949	164	25	ElasticNet	RoC maximization	10*Repeated 10-fold CV	NIMH [2]	yes
Chekroud et al. (2017)	Antidepressant treatment	7221	164	25	ElasticNet	NR	5-fold CV	NIMH [2]	yes
Andrews et al. (2017)	Depression in older adults	37	6	2	LASSO	Stratified CV	5-fold CV	NR	no
Nijeweme-d'Hollosy et al. (2018)	LBP self-referral	1288	15	NR	Algorithm implicit	NR	Holdout (70,30)	On Request	yes
Rahman et al. (2018)	Pain volatility	782	130	NR	Algorithm implicit	NR	5-fold CV	NR	no
Rahman et al. (2019)	Pain volatility	879	132	9	Gini impurity, Information gain, Class imbalance	NR	5-fold CV	NR	no
Schiltz et al. (2020)	Hospital Readmission	6617	NR	NR	RF	NR	Holdout (80,20)	HRS [3]	no
Wang et al. (2020b)	Oral Health	908	27	NA	Manual	Greedy approximation	Holdout (70,30)	NR	yes

[1] *www.digital.nhs.uk/data*; [2] *www.nimhgenetics.org/download−tool/DP*; [3] *www.hrsonline.isr.umich.edu*.

(Pua et al., 2019) and previous experimental evaluation (Harris et al., 2019). Ten articles employed algorithmic approach for extraction and selection of relevant features from the datasets. Andrews et al. (2017) used LASSO, Schiltz et al. (2020) and Rahman et al. (2019) used Random Forest, Polce et al. (2020) used recursive feature elimination with Random Forest, Chekroud et al. (2016, 2017) used Elastic nets, while Huber et al. (2019), Rahman et al. (2018), Nijeweme-d'Hollosy et al. (2018) and Kessler et al. (2016) employed no separate feature selection but relied on the implicit feature selection ability of the algorithms used. Random forest and linear models such as Elastic nets and LASSO appear to be the preferred algorithm choice for feature selection.

B.VIII Trends in the Application of Machine Learning Methods

Table 2 presents an overview of the different machine learning methods used in the included articles. Ensembles and linear methods appear to be the most commonly applied methods to PROMs datasets, with all the included articles employing at least either one, likely due to their ability to extract features implicitly. While supervised learning methods are the go-to methods for prediction tasks, three (20%) articles apply unsupervised methods as a pre-step to the supervised methods to determine and predict cluster-specific outcomes (Rahman et al., 2018, 2019; Chekroud et al., 2017). Examples of commonly used linear algorithms in the included articles are logistic regression, logistic regression with splines, elastic nets, Poisson regression, LASSO, linear kernel-based Support Vector Machines, among others. The most commonly applied ensemble algorithms are Random Forest, Boosted Trees, Gradient Boosting Machines (GBM), stochastic gradient boosting machines, extreme gradient boosting (XGBoost), and SuperLearner.

Thirteen (87%) articles used binary classification to predict whether the targeted outcome(s) are above or below a specified threshold (for instance, whether or not a patient achieves MCID in their post-operative outcomes (Fontana et al., 2019)). One article used ternary classification to predict the self-referral outcome among people with low back pain in a primary care setting (Nijeweme-d'Hollosy et al., 2018). In contrast, three (20%) articles used regression (Shi et al., 2012; Chekroud et al., 2017; Wang et al., 2020b), one of which used both regression and binary classification to predict continuous and categorical outcomes (Wang et al., 2020b).

Table 2: Overview of the application of different machine learning methods in the included articles. Abbreviations- EM: Ensemble Methods, LM: Linear Methods, DT: Decision Tree, SVM: Support Vector Machines, NN: Neural Network, NB: Naive Bayes, k-NN: k-Nearest Neighbour, QDA: Quadratic Discriminant Analysis, Aggl: Agglomerative Clustering, ML: Machine Learning.

Article	**Supervised**							**Unsupervised**		**ML Task**
	EM	**LM**	**DT**	**SVM**	**NN**	**NB**	**k-NN**	**QDA**	**k-Means**	**Aggl.**	
Shi et al. (2012)		✓			✓						Regression
Huber et al. (2019)	✓	✓			✓	✓	✓				Classification
Fontana et al. (2019)	✓	✓		✓							Classification
Polce et al. (2020)	✓	✓		✓	✓						Classification
Pua et al. (2019)	✓	✓									Classification
Harris et al. (2019)	✓	✓						✓			Classification
Kessler et al. (2016)	✓	✓									Classification
Chekroud et al. (2016)	✓										Classification
Chekroud et al. (2017)	✓									✓	Regression
Andrews et al. (2017)		✓									Classification
Nijeweme-d'Hollosy et al. (2018)	✓		✓								Classification
Rahman et al. (2018)	✓	✓		✓					✓		Classification
Rahman et al. (2019)	✓	✓		✓					✓		Classification
Schiltz et al. (2020)	✓	✓	✓								Classification
Wang et al. (2020b)	✓					✓					Regression, Classification

B.IX Study Design and Model Evaluation

To reduce the risk of overfitting the models and improve their generalizability, k-fold cross-validation scheme was used in eleven articles, either during the hyperparameter tuning phase or the model evaluation phase (Table 1). Out of these eleven, only one article used k-fold cross-validation scheme in both phases (Andrews et al., 2017). Three articles (Wang et al., 2020b; Nijeweme-d'Hollosy et al., 2018; Pua et al., 2019) employed a holdout (70,30) validation approach: 70% of the dataset used for training the model and 30% for validation, while four articles employed a holdout (80,20) validation approach (Schiltz et al., 2020; Polce et al., 2020; Fontana et al., 2019; Shi et al., 2012). While the holdout validation approach is useful due to its speed and simplicity, it often leads to high variability due to the differences in the training and test datasets, which can result in significant differences in the evaluation metric estimates (accuracy, error, sensitivity etc, depending on the machine learning task the metric used).

External validation datasets were used in four articles to test the generalizability of the models (Wang et al., 2020b; Nijeweme-d'Hollosy et al., 2018; Chekroud et al., 2016, 2017). While external validation is generally recommended to validate the models generated since prediction models perform better on the training data than on new data, internal validation appears to be more common, likely due to either lack of or unavailability of appropriate external validation dataset. However, to correct the bias in the internally-validated prediction models, bootstrapping methods are recommended (Bleeker et al., 2003; Steyerberg and Harrell Jr, 2016). Only one article used bootstrapping to internally validate the models where external validation dataset was not used (Harris et al., 2019).

B.X Model Performance

While it's difficult to provide a concrete result comparison among the included articles due to utilisation of various metrics, most articles did report at least above chance (fair to moderate) predictive performance of the machine learning models. Amongst the articles that compared the performance of conventional linear models with machine learning models, most found the machine learning models to perform better for predicting the outcomes (Shi et al., 2012; Huber et al., 2019; Kessler et al., 2016), while one article found the conventional method to perform equally well as the machine learning

methods (Pua et al., 2019). Despite the above chance predictive performance reported in most articles, the limitations posed by the small size of training datasets used to develop the models and the lack of external validation datasets has been widely acknowledged (Wang et al., 2020b; Shi et al., 2012; Polce et al., 2020; Andrews et al., 2017).

C Discussion

Our review identified 15 articles focusing on utilization of PROMs for predicting outcomes by leveraging the analytical abilities of machine learning methods. Over the last decade, machine learning methods have received more attention in clinical research and are increasingly being adopted for furthering research in clinical analysis, modeling and building decision support systems for practitioners. The included articles presented promising research, demonstrating that as more and more healthcare data becomes available for developmental research, personalized treatment and medicine becomes more feasible with the help of machine learning-based decision support systems. Mobile applications allowing faster collection of PROMs data, as shown by (Rahman et al., 2018, 2019), is a promising way to collect more data frequently as well as utilise the collected data for further research and development. Thus, the application of machine learning methods on PROMs data for predicting patient-specific outcomes appears to be a promising avenue and warrants further research.

C.I Gaps and Challenges

The lack of external validation and non-availability of datasets used in the majority of the articles poses a major gap in data availability for machine learning research. To drive the field forward, access to and open research questions in suitable datasets is a prerequisite. Datasets that are both comprehensive, complete, and readily available for research purposes such as machine learning model development are rare. Such datasets can facilitate the external validation by researchers in different disciplines and potentially inter-disciplinary collaboration. In other medical domains, opening pre-processed and experiment-ready datasets have shown that they draw attention to machine learning researchers and practitioners to explore different methods and benchmark the results (Xu et al., 2019; Wang et al.,

2020a; Feng et al., 2018). As for the size of the datasets, eight of the fifteen articles included in this review used training datasets with more than 1000 patients (see Table 1), highlighting the sparsity of decent sized healthcare datasets for machine learning modelling. Furthermore, data collected with a different intent originally cannot automatically be used for machine learning due to uncertain or missing informed consents from participants. Most datasets collected from patients requires their consent for utilisation of their data for various other purposes which may not have been foreseen at the time of data collection. This may limit the ways patient data can be stored, used or distributed as well as the scope of the data.

Explainability and trustworthiness of the machine learning models are important challenges when it comes to developing clinical decision support systems. While a lot of attention has been given to developing accurate machine learning models, it is crucial to build systems that are trustworthy and interpretable. The users of such systems, for example medical researchers or clinicians, should be able to interpret the output of the machine learning models. Interpretations can be facilitated either through visualizations or explanations. This is an important aspect for clinicians such that they can focus on addressing the medical concerns rather than struggling with comprehension of the system's results.

Moreover, inconsistency was observed in reporting the development of the machine learning models in the articles. Only six articles reported the essential aspects of machine learning model development such, as feature selection and hyperparameter tuning, whereas in nine articles this was either unclear or not stated at all, which can limit the reproducibility of results and further research.

Despite the progress in the development of machine learning models aimed at facilitating informed decision-making, there is still some more progress needed before these tools can be used in clinical practice. In specific, external validation on large datasets of specific cohorts and thorough evaluation of the prediction tools would be necessary before these tools can be integrated in clinical practice.

C.II Limitations

The limitations of this review were that it was not possible to perform a meta-analysis of the results in the included articles due to various reasons, including, but not limited to, the heterogeneous study design, data non-

availability, and study results, as summarised in Table 1 and discussed in Section B.X. Out of the fifteen articles included in the analysis, only four articles reported their data source (national registry datasets) and one article stated that their dataset may be available on a reasonable request. However, none of the datasets were available during this literature review process for a meta-analysis. Further, we acknowledge that the articles retrieved in this literature review include only those articles that were retrieved during our search and met the inclusion criteria. As stated in the inclusion criteria, we included only those articles that focus solely on PROMs.

D Conclusions

In summary, this literature review resulted in two main findings. First, there has been an increase during recent years in applying machine learning methods in exploring PROMs datasets for predicting patient-specific outcomes. Second, although the included articles have reported promising results and improvements (Chekroud et al., 2016; Shi et al., 2012; Pua et al., 2019), lack of data availability, inconsistent reporting of machine learning model development as well as the use of different evaluation metrics prevents effective result reproduction and comparison. To conclude, utilising machine learning methods on PROMs datasets have the potential for assisting in clinical decision making and thereby, further research focusing on thorough validation is needed.

Bibliography

JA Andrews, RF Harrison, LJE Brown, LM MacLean, F Hwang, T Smith, Elizabeth A Williams, C Timon, T Adlam, H Khadra, et al. Using the nana toolkit at home to predict older adults' future depression. *Journal of affective disorders*, 213:187–190, 2017.

Michael J Barry and Susan Edgman-Levitan. Shared decision making—the pinnacle patient-centered care. *New England Journal of Medicine*, 366(9): 780–781, 2012. doi: 10.1056/NEJMp1109283. PMID: 22375967.

Judith Baumhauer. Patient-reported outcomes- are they living up to their potential? *New England Journal of Medicine*, 377:6–9, 07 2017. doi: 10.1056/NEJMp1702978.

Clifton O Bingham III, Vanessa K Noonan, Claudine Auger, Debbie E Feldman, Sara Ahmed, and Susan J Bartlett. Montreal accord on patient-reported outcomes (pros) use series–paper 4: patient-reported outcomes can inform clinical decision making in chronic care. *Journal of clinical epidemiology*, 89:136–141, 2017.

SE Bleeker, HA Moll, EW Steyerberg, ART Donders, Gerarda Derksen-Lubsen, DE Grobbee, and KGM Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832, 2003.

Adam M Chekroud, Ralitza Gueorguieva, Harlan M Krumholz, Madhukar H Trivedi, John H Krystal, and Gregory McCarthy. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry*, 74(4):370–378, 2017.

Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon,

John Harrison Krystal, and Philip Robert Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, 3(3):243–250, 2016.

Valle Coronado-Vázquez, Carlota Canet-Fajas, Maria Teresa Delgado-Marroquín, Rosa Magallón-Botaya, Macarena Romero-Martín, and Juan Gómez-Salgado. Interventions to facilitate shared decision-making using decision aids with patients in primary health care: A systematic review. *Medicine*, 99(32), 2020.

S Prasanna Devi, K Suryaprakasa Rao, and S Sai Sangeetha. Prediction of surgery times and scheduling of operation theaters in optholmology department. *Journal of medical systems*, 36(2):415–430, 2012.

Mengling Feng, Jakob I McSparron, Dang Trung Kien, David J Stone, David H Roberts, Richard M Schwartzstein, Antoine Vieillard-Baron, and Leo Anthony Celi. Transthoracic echocardiography and mortality in sepsis: analysis of the mimic-iii database. *Intensive care medicine*, 44(6):884–892, 2018.

Mark Alan Fontana, Stephen Lyman, Gourab K Sarker, Douglas E Padgett, and Catherine H MacLean. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research*, 477 (6):1267–1279, 2019.

Aliyah Giga. How health leaders can benefit from predictive analytics. In *Healthcare management forum*, volume 30, pages 274–277. SAGE Publications Sage CA: Los Angeles, CA, 2017.

Alex HS Harris, Alfred C Kuo, Yingjie Weng, Amber W Trickey, Thomas Bowe, and Nicholas J Giori. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research*, 477(2):452, 2019.

Manuel Huber, Christoph Kurz, and Reiner Leidl. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC medical informatics and decision making*, 19(1):3, 2019.

Prakash Jayakumar, Jianing Di, Jiayu Fu, Joyce Craig, Vicki Joughin, Victoria Nadarajah, Jade Cope, Marcus Bankes, Peter Earnshaw, and Zameer Shah. A patient-focused technology-enabled program improves outcomes in primary total hip and knee replacement surgery. *JBJS Open Access*, 2(3), 2017.

Zahra Shahabi Kargar, Sankalp Khanna, and Abdul Sattar. Using prediction to improve elective surgery scheduling. *The Australasian medical journal*, 6(5):287, 2013.

Ronald C Kessler, Hanna M van Loo, Klaas J Wardenaar, Robert M Bossarte, Lisa A Brenner, Tianxi Cai, David Daniel Ebert, Irving Hwang, Junlong Li, Peter de Jonge, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular psychiatry*, 21(10):1366–1371, 2016.

Charlotte Kingsley and Sanjiv Patel. Patient-reported outcome measures and patient-reported experience measures. *Bja Education*, 17(4):137–144, 2017.

M Krismer, M Van Tulder, et al. Low back pain (non-specific). *Best practice & research clinical rheumatology*, 21(1):77–91, 2007.

Gemma Mansell, Nadia Corp, Gwenllian Wynne-Jones, Jonathan Hill, Siobhán Stynes, and Daniëlle van der Windt. Self-reported prognostic factors in adults reporting neck or low back pain: An umbrella review. *European Journal of Pain*, 2021.

Takawira C Marufu, SM White, R Griffiths, SR Moonesinghe, and Iain K Moppett. Prediction of 30-day mortality after hip fracture surgery by the nottingham hip fracture score and the surgical outcome risk tool. *Anaesthesia*, 71(5):515–521, 2016.

Suneetha Ramani Moonesinghe, Michael G Mythen, Priya Das, Kathryn M Rowan, and Michael PW Grocott. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology*, 119(4):959–981, 2013.

Wendy Oude Nijeweme-d'Hollosy, Lex van Velsen, Mannes Poel, Catharina G. M. Groothuis-Oudshoorn, Remko Soer, and Hermie Hermens. Evaluation

of three machine learning models for self-referral decision support on low back pain in primary care. *International Journal of Medical Informatics*, 110:31–41, 2 2018. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2017.11.010.

Evan M. Polce, Kyle N. Kunze, Michael Fu, Grant E. Garrigues, Brian Forsythe, Gregory P. Nicholson, Brian J. Cole, and Nikhil N. Verma. Development of supervised machine learning algorithms for prediction of satisfaction at two years following total shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery*, 2020. ISSN 1058-2746. doi: https://doi.org/10.1016/j.jse.2020.09.007.

Yong-Hao Pua, Hakmook Kang, Julian Thumboo, Ross Allan Clark, Eleanor Shu-Xian Chew, Cheryl Lian-Li Poon, Hwei-Chi Chong, and Seng-Jin Yeo. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. *Knee Surgery, Sports Traumatology, Arthroscopy*, pages 1–10, 2019.

Quazi Abidur Rahman, Tahir Janmohamed, Meysam Pirbaglou, Hance Clarke, Paul Ritvo, Jane M Heffernan, and Joel Katz. Defining and predicting pain volatility in users of the manage my pain app: Analysis using data mining and machine learning methods. *Journal of medical Internet research*, 20(11):e12001, 2018.

Quazi Abidur Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, and Joel Katz. Interpretability and class imbalance in prediction models for pain volatility in manage my pain app users: analysis using feature selection and majority voting methods. *JMIR medical informatics*, 7(4):e15601, 2019.

Nicholas Schiltz, Mary Dolansky, David Warner, Kurt Stange, Stefan Gravenstein, and Siran Koroukian. Impact of instrumental activities of daily living limitations on hospital readmission: an observational study using machine learning. *Journal of General Internal Medicine*, 07 2020. doi: 10.1007/s11606-020-05982-0.

Karen R. Sepucha, Steven J. Atlas, Yuchiao Chang, Andrew Freiberg, Henrik Malchau, Mahima Mangla, Harry Rubash, Leigh H. Simmons, and Thomas Cha. Informed, patient-centered decisions associated with better health outcomes in orthopedics: Prospective cohort study. *Medical Decision Making*, 38(8):1018–1026, 2018. doi: 10.1177/0272989X18801308.

Zahra ShahabiKargar, Sankalp Khanna, Norm Good, Abdul Sattar, James Lind, and John O'Dwyer. Predicting procedure duration to improve scheduling of elective surgery. In Duc-Nghia Pham and Seong-Bae Park, editors, *PRICAI 2014: Trends in Artificial Intelligence*, pages 998–1009, Cham, 2014. Springer International Publishing. ISBN 978-3-319-13560-1.

Hon-Yi Shi, Jinn-Tsong Tsai, Yao-Mei Chen, Richard Culbertson, Hong-Tai Chang, and Ming-Feng Hou. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. *Breast cancer research and treatment*, 135(1):221–229, 2012.

Amit G Singal, Ashin Mukherjee, B Joseph Elmunzer, Peter DR Higgins, Anna S Lok, Ji Zhu, Jorge A Marrero, and Akbar K Waljee. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *The American journal of gastroenterology*, 108(11):1723, 2013.

Ewout W Steyerberg and Frank E Harrell Jr. Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, 69:245, 2016.

Akbar K Waljee, Peter DR Higgins, and Amit G Singal. A primer on predictive models. *Clinical and translational gastroenterology*, 5(1):e44, 2014.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 222–235, 2020a.

X. Wang and V. Gottumukkala. Patient reported outcomes: Is this the missing link in patient-centered perioperative care? *Best Practice & Research Clinical Anaesthesiology*, 2020.

Yan Wang, Ronald D Hays, Marvin Marcus, CA Maida, J Shen, D Xiong, ID Coulter, SY Lee, VW Spolsky, JJ Crall, et al. Developing children's oral health assessment toolkits using machine learning algorithm. *JDR Clinical & Translational Research*, 5(3):233–243, 2020b.

DJN Wong, CM Oliver, and SR Moonesinghe. Predicting postoperative morbidity in adult elective surgical patients using the surgical outcome risk tool (sort). *BJA: British Journal of Anaesthesia*, 119(1):95–105, 2017.

Jinghong Xu, Li Tong, Jiyou Yao, Zilu Guo, Ka Yin Lui, XiaoGuang Hu, Lu Cao, Yanping Zhu, Fa Huang, Xiangdong Guan, et al. Association of sex with clinical outcome in critically ill sepsis patients: a retrospective analysis of the large clinical database mimic-iii. *Shock (Augusta, Ga.)*, 52 (2):146, 2019.

External Validation of Prediction Models for Patient-Reported Outcome Measurements collected using the SELFBACK Mobile App

Deepika Verma, Kerstin Bach, Paul Jarle Mork

Abstract

External validation is essential in examining the disparities in the training and validation cohorts during the development of prediction models, especially when the application domain is healthcare-oriented. Currently, the use of prediction models in healthcare research aimed at utilising the under-explored potential of patient-reported outcome measurements (PROMs) is limited, and few are validated using external datasets. To address the issue, we validate the machine learning prediction models based on three methods—Case-Based Reasoning, Support Vector Regression, and XGBoost Regression—developed in our previous work (Verma et al., 2021a) for predicting pain-related patient-reported outcomes from the SELFBACK datasets. Overall, the predictive power was low, except for prediction of one of the outcomes. The results indicate that the models show ability to generalise and predict outcomes for a new dataset and highlight the need for external validation in healthcare-oriented studies for the further development of patient-centred healthcare systems.

A Introduction

Use of technology to support self-management of musculoskeletal pain is a feasible and promising approach (Marcuzzi et al., 2021; Sandal et al., 2019). In the SELFBACK project, a mobile app was developed to make weekly tailored self-management plans for users to help them manage back pain and other pain-related symptoms (Mork and Bach, 2018). The self-management plans are tailored to each user based on a set of variables reported by the user in the mobile application. Tools like SELFBACK enable the effective use of technology for bridging the gap between patient-reported outcome measurements (PROMs) and patient-centred care. PROMs serve as a tool to assess and evaluate the health status of a patient from the patient's perspective at any given time point (Nelson et al., 2015). They may be recorded before, during or after a healthcare intervention and can help in measuring the impact of the intervention given to the patient. From a clinical perspective, the addition of predictive analytics to such healthcare tools could serve to further improve patient-centred care by detecting early signs of deteriorating outcomes, and warning primary caregivers to proactively prevent their occurrence (White et al., 2020; Verma et al., 2021b). This can therefore help caregivers to optimise the treatment approach for a given patient.

Previous research has shown that integrating technology with healthcare data can support preventive treatment (Chekroud et al., 2016; Andrews et al., 2017), hospital re-admissions (Schiltz et al., 2020), and prevention of post-surgical complications

(Harris et al., 2019). To make further advancements in this field, it is important to have a clear understanding of what factors should be considered when deciding the treatment approach for a given patient (Giga, 2017). From both clinical and machine learning points of view, this translates to deciding features from the available data that may be valuable in predicting a future outcome. Furthermore, external validation is essential to assess the generalisability of the prediction models (Moons et al., 2006; Cabitza et al., 2021).

Most studies that address the prediction of PROMs using machine learning methods have only validated their models internally (Verma et al., 2021c). Bootstrapping may be an approach suitable for internal validation to compensate for the lack of external validation due to the bias-corrected estimation of the prediction models (Steyerberg and Harrell, 2016). However, the bootstrapping method cannot replace external validation since models often perform better on the dataset they were trained on compared to validation on a different dataset (Bleeker et al., 2003). This effect is often attributed to the overfitting of the model caused by the high variance. Furthermore, since clinical datasets tend to be relatively small, it is unlikely that internal validation would be sufficient as prediction models are prone to overfitting when using small datasets (Luedtke et al., 2019).

This paper presents an evaluation of the prediction models developed in our previous work (Verma et al., 2021a) using an external dataset. In our previous work, a two-fold feature selection approach that combines correlation and data-driven similarities in Case-Based Reasoning (CBR) was used to identify relevant features for predicting a set of PROMs in the SELFBACK dataset. The features selected were used to build prediction models using three methods—CBR, Support Vector Regression (SVR), and XGBoost Regression (XGB).

B Methods

B.I Dataset

The dataset used for external validation consists of PROMs collected from patients with non-specific neck and/or low back pain in a randomised controlled trial (RCT II[1]) with the help of questionnaires to evaluate the effectiveness of the SELFBACK decision support system (DSS) in a secondary care setting (Marcuzzi et al., 2021). The dataset used for training the models consisted of PROMs collected from patients with

[1] https://clinicaltrials.gov/ct2/show/NCT04463043

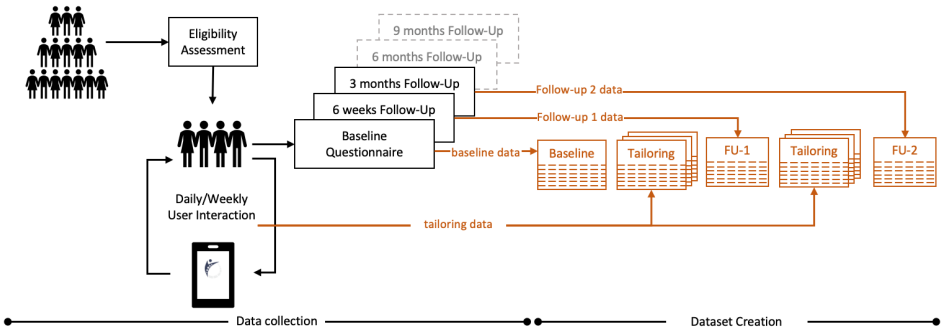


Figure 1: Overview of data collection in two RCTs that evaluated the SELFBACK DSS. The different data components are indicated by the orange boxes. Only data from baseline and the 3-month follow-up (FU-2 data) is used to train and evaluate the prediction models.

non-specific low back pain during an earlier RCT (I^2) with the help of questionnaires to evaluate the SELFBACK DSS in a primary care setting (Sandal et al., 2019).

Figure 1 shows how the data collection was carried out in the two RCTs. The collected data is categorised into Baseline, Tailoring, and Follow-Up (FU). Only data from baseline and the 3-month follow-up 2 (FU-2 data) is used to train and evaluate the prediction models. In total, the training dataset includes 218 patients while the external validation dataset includes 75 patients that completed at least the FU-2 questionnaire. The external validation dataset is a subset of the data collected in RCT II. A detailed account of the data collection in the two RCTs can be found in Sandal et al. (2019) and Marcuzzi et al. (2021).

During data collection (for the external validation dataset), eligible patients who accepted to join the study answered questionnaires at different time points: (1) at the time of intake: Baseline questionnaire (*Baseline Data*), (2) at the end of every week: Tailoring questionnaire (*Tailoring Data*), (3) at the end of 6-weeks, 3-months, 6-months: Follow-Up questionnaire ((*FU Data*)). The questionnaires include validated clinical measures of pain level, pain self-efficacy, work-ability, mood, physical activity, sleep quality, functional ability, and fear avoidance. In addition to the clinical measures, the baseline questionnaire also includes questions regarding patient demographics such as age, height, weight, education, employment type, and family. Based on the patients' responses at baseline, the SELFBACK mobile application recommends an exercise plan and educational elements along

[2] https://clinicaltrials.gov/ct2/show/NCT03798288

with tracking their number of steps every day from a wearable device (Xiaomi Mi Band 3). Exercise completion and education readings are self-reported in the app (Sandal et al., 2021).

Target Outcomes
The training dataset originally comprised 47 features. In our previous work, we focused on six target outcomes. However, due to exclusion of one outcome in RCT II, two outcomes—Roland Morris Disability Questionnaire and Numeric Pain Rating Scale—had to be removed from this experimental evaluation due to feature dependency. Instead, we focus on the four secondary outcomes that were chosen to represent a diversity of domains; Workability index (WAI, range: [0,10]), Pain Self Efficacy Questionnaire (PSEQ, range: [0,60]), Fear Avoidance Belief Questionnaire (FABQ, range: [0,30]) and Global Perceived Effect Scale (GPE, range: [-5,+5]). We use the features previously selected for each target outcome based on the training dataset (Verma et al., 2021a) and evaluate the generalisability of the models using the external dataset. Table 1 gives a brief summary of the various features used in this work. Marcuzzi et al. (2021) give a more comprehensive summary of all the features collected at various time points in the RCT II.

B.II Prediction Models

Prediction models using three machine learning methods were trained on the completed PROMs collected in RCT I—CBR, SVR, and XGB—to predict the four target outcomes reported by patients in RCT II.

Table 1: Summary of the SELFBACK RCT I & II dataset features used in this work. Abbreviated feature names in the bracket include the specific sub-scale scores used in this work either as a predictor or as a target outcome. Features predicted at FU-2 are marked with an asterisk (*).

Feature	**Description**
Age	Age of the participant in years
Body Mass Index (BMI)	Calculated using reported weight and height

(continued on next page)

Feature	Description
Workability Index (WAI*)	Used to assess work-ability of an individual using an 11-point numeric rating scale
Pain Self-Efficacy Questionnaire (PSEQ*, PSEQ_2)	Used to assess the participants' level of confidence in carrying out specific activities despite their pain using ten items, each measured on a 6-point scale
EuroQoL 5-dimension (EQ5D, EQ5D_mobility)	Used to assess health-related quality of life using five items, each scored 0-5
Brief Illness Perception Questionnaire (BIPQ_life, BIPQ_pain_continuation, BIPQ_concern, BIPQ_symptoms)	Used to evaluate participants' illness perception using eight items on an 11-point numeric rating scale
Pain Intensity (Pain_1year, Pain_worst)	Perceived intensity of low back and/or neck pain measured by a 11-point numerical rating scale
Sleep (Sleep_wakeup)	Sleep problems assessed by four self-report items which provide information needed to diagnose insomnia according to the DSM-V criteria
Fear-Avoidance Belief Questionnaire (FABQ*)	Physical activity sub-scale used to measure participant's beliefs about how physical activity affects their low back and/or neck pain using five items, each scored 0-6
Patient-Specific Functional Scale (PSFS)	Used to evaluate changes in participant's ability to perform up to two self-selected activities regarded as important by them using an 11-point score
Global Perceived Effect (GPE*)	Used to investigate the effect of the intervention as perceived by the participant using one item scored -5 to 5

B.III Feature Selection

Two feature selection approaches were applied in the previous work to select features from the baseline dataset for each of the chosen outcomes to be predicted: i) a two-fold hybrid approach that uses statistical correlation (Pearson and Galton, 1895) to filter out the most correlated features, followed by a final selection of features based on CBR model built using data-driven local similarity modelling approach (Verma et al., 2018) (similarity modelling carried out in *myCBR workbench* (Bach and Althoff, 2012)) and ii) an ensemble approach that uses permutation feature importance to select features with XGBoost as the base regressor (Fisher et al., 2019).

B.IV Hyperparameter Optimization

Before the two machine learning algorithms—SVR and XGB—were trained on the training dataset, their hyperparameters were tuned using grid search to optimize their performance on the dataset. Grid search was used to perform an exhaustive search through a pre-defined set of hyperparameter space for each learning algorithm to identify their optimal hyperparameters (Hutter et al., 2019). Regarding the CBR models, as there are no hyperparameters involved, this step was not required.

B.V Evaluation Metrics

The metrics used to evaluate the results in the experiments are Mean Absolute Error (MAE) and Normalized Mean Absolute Error (NMAE). MAE is the average of the absolute errors, i.e., the difference between the observed and predicted value. While there are several ways to normalize error, we normalized the MAE using the max-min method (see eq.2) for each outcome to get NMAE in the range [0,1]. This brings the results on the same scale and simplifies comparison across different models and outcomes.

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{1}$$

$$NMAE = \left(\frac{MAE}{y_{max} - y_{min}}\right) \tag{2}$$

C Experiments & Results

The methods were implemented in Python (Oliphant, 2007) in jupyterLab notebook[3] using Scikit-learn (Buitinck et al., 2013) and *myCBR* Rest API[4] (Bach et al., 2019) was used for querying the CBR models developed in *myCBR workbench*. SVR and XGB models were 10-fold cross-validated during the training phase.

Table 2: Results of Prediction of Target Outcomes at Followup-2 using different Feature Selection Methodologies and Regression Methods for the Intervention Group (size of dataset: 75 participants). Values are MAE/NMAE pairs. Numbers in bold letters highlight lowest MAE/NMAE pair. **n**: number of features.

	Feature Selection Methodology							
	Correlation+CBR				**PFI+XGBoost**			
Outcome [range]	**n**	**CBR**	**SVR**	**XGB**	**n**	**CBR**	**SVR**	**XGB**
WAI [0,10]	4	2.04/0.204	**1.68/0.168**	1.94/0.194	1	1.90/0.190	1.91/0.191	1.92/0.192
PSEQ [0,60]	3	9.97/0.166	9.49/0.158	8.66/0.144	2	10.28/0.171	9.8/0.163	**8.04/0.134**
FABQ [0,30]	1	5.54/0.184	4.09/0.133	**4.04/0.134**	6	5.24/0.174	4.34/0.144	4.74/0.158
GPE [-5,5]	2	1.32/0.132	1.92/0.192	1.55/0.155	3	**1.30/0.130**	1.60/0.160	1.43/0.143

Table 2 summarises the results of the experiments. The SVR and XGB models gave the lowest prediction error for *WAI* and *FABQ* at *1.68* and *4.04*, respectively, using the features selected by the hybrid method. XGB and CBR gave the lowest prediction error for *PSEQ* and *GPE* at *8.04* and *1.30*, respectively, using the features selected by the permutation feature importance method. For the MAE, the error for *PSEQ* and *FABQ* is higher compared to the other two outcomes, however, considering the NMAE, these errors are comparable to that of the other two outcomes.

D Discussion

The external validation sample in this work included PROMs from 75 patients, while the training and internal validation included 218 participants. In the internal

[3] `https://jupyter.org/`

[4] `https://github.com/ntnu-ai-lab/mycbr-sdk`

validation of the models in our previous work (Verma et al., 2021a), CBR and SVR gave the lowest MAE for *WAI* and *FABQ* at *1.14* and *3.60*, respectively, using the features selected by the hybrid method, while CBR gave the lowest MAE for *PSEQ* and *GPE* at *5.95* and *1.49*, respectively, using the features selected by the feature importance method. Comparing these figures to the results in table 2, we can see that the models show slightly worse performance for the external dataset, which is usually expected. While the results for *PSEQ* appear worse in the external validation, when considering the performance of the same best-performing model in the internal validation for the outcome (XGB), the model in fact fared better on the external dataset (MAE 8.04) that on the training dataset (MAE 17.1). Although the predictive power was low, the evaluation suggest that the prediction models can be applied to a new dataset. The approach for selecting features seems to have negligible influence on the performance of the prediction models.

Training and testing a predictive model on the same dataset is by and large not considered optimal, especially when the predictions should be used to support clinical decision-making (Siontis et al., 2015). At a minimum, our evaluation substantiates the potential of both the PROMs and utility of machine learning methods for PROMs, while also highlighting the need for external validation and further development of prediction models. Future work should compare the predictions made by clinicians versus machine learning methods to fully assess the usefulness of machine learning methods in this field.

D.I Study Limitation

The fact that this work is based on patient-reported data may be considered a limitation owing to the limited reliability of subjective datasets (Bookstein and Lindsay, 1989). Further, it is difficult to fully assess the extent of adequacy of the features selected for clinical judgement since clinicians themselves have a hard time selecting the most valuable or informative features (Leuchter et al., 2009).

E Conclusion

To conclude, the external validation of prediction models presents modest results and highlights the need for further development in this area of machine learning application. While the results are still far from being applicable in a clinical setting, they nevertheless show potential in the methods as well as PROMs data. More research is prudent to further this field of machine learning application.

Bibliography

JA Andrews, RF Harrison, LJE Brown, LM MacLean, F Hwang, T Smith, Elizabeth A Williams, C Timon, T Adlam, H Khadra, et al. Using the nana toolkit at home to predict older adults' future depression. *Journal of affective disorders*, 213:187–190, 2017.

Kerstin Bach and Klaus-Dieter Althoff. Developing Case-Based Reasoning Applications Using myCBR 3. In Ian Watson and Belen Diaz Agudo, editors, *Case-based Reasoning in Research and Development, Proceedings of the 20th International Conference on Case-Based Reasoning (ICCBR-12)*, pages 17–31. LNAI 6880, Springer, 2012.

Kerstin Bach, Bjørn Magnus Mathisen, and Amar Jaiswal. Demonstrating the mycbr rest api. In *ICCBR Workshops*, pages 144–155, 2019.

SE Bleeker, HA Moll, EWet al Steyerberg, ART Donders, Gerarda Derksen-Lubsen, DE Grobbee, and KGM Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832, 2003.

Abraham Bookstein and A Lindsay. Questionnaire ambiguity: A rasch scaling model analysis. *Graduate School of Library and Information Science. University of Illinois . . .* , 1989.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Federico Cabitza, Andrea Campagner, Felipe Soares, Luis GarcÃaÂ deÂ Guadiana-Romualdo, Feyissa Challa, Adela Sulejmani, Michela Seghezzi, and Anna Carobene. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, 208:106288, 2021. ISSN 0169-2607.

Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon, John Harrison Krystal, and Philip Robert Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, 3(3):243–250, 2016.

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. *Journal of Machine Learning Research*, 20(177): 1–81, 2019.

Aliyah Giga. How health leaders can benefit from predictive analytics. In *Healthcare management forum*, volume 30, pages 274–277. SAGE Publications Sage CA: Los Angeles, CA, 2017.

Alex HS Harris, Alfred C Kuo, Yingjie Weng, Amber W Trickey, Thomas Bowe, and Nicholas J Giori. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research*, 477(2):452, 2019.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning*. Springer, 2019.

Andrew F Leuchter, Ian A Cook, Lauren B Marangell, William S Gilmer, Karl S Burgoyne, Robert H Howland, Madhukar H Trivedi, Sidney Zisook, Rakesh Jain, James T McCracken, Maurizio Fava, Dan Iosifescu, and Scott Greenwald. Comparative effectiveness of biomarkers and clinical indicators for predicting outcomes of ssri treatment in major depressive disorder: results of the brite-md study. *Psychiatry research*, 169(2), September 2009. ISSN 0165-1781. doi: 10.1016/j.psychres.2009.06.004.

Alex Luedtke, Ekaterina Sadikova, and Ronald C Kessler. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, 7(3):445–461, 2019.

Anna Marcuzzi, Kerstin Bach, Anne Lovise Nordstoga, Gro Falkener Bertheussen, Ilya Ashikhmin, Nora Østbø Boldermo, Else-Norun Kvarner, Tom Ivar Lund Nilsen, Gunn Hege Marchand, Solveig Osborg Ose, and et al. Individually tailored self-management app-based intervention (selfback) versus a self-management web-based intervention (e-help) or usual care in people with low back and neck pain referred to secondary care: protocol for a multiarm randomised clinical trial. *BMJ Open*, 11(9), 2021. doi: 10.1136/bmjopen-2020-047921.

Karel GM Moons, Rogier ART Donders, Theo Stijnen, and Frank E Harrell Jr. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, 59(10):1092–1101, 2006.

Paul Jarle Mork and Kerstin Bach. A decision support system to enhance self-management of low back pain: protocol for the selfback project. *JMIR research protocols*, 7(7):e167, 2018.

Eugene C Nelson, Elena Eftimovska, Cristin Lind, Andreas Hager, John H Wasson, and Staffan Lindblad. Patient reported outcome measures in practice. *Bmj*, 350, 2015.

Travis E Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

Karl Pearson and Francis Galton. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

Louise Fleng Sandal, Mette Jensen Stochkendahl, Malene Jagd Svendsen, Karen Wood, Cecilie K Øverås, Anne Lovise Nordstoga, Morten Villumsen, Charlotte Diana Nørregaard Rasmussen, Barbara Nicholl, Kay Cooper, Per Kjaer, Frances S Mair, Gisela Sjøgaard, Tom Ivar Lund Nilsen, Jan Hartvigsen, Kerstin Bach, Paul Jarle Mork, and Karen Søgaard. An app-delivered self-management program for people with low back pain: Protocol for the selfback randomized controlled trial. *JMIR Res Protoc*, 8(12):e14720, Dec 2019. ISSN 1929-0748. doi: 10.2196/14720.

Louise Fleng Sandal, Kerstin Bach, Cecilie K Øverås, Malene Jagd Svendsen, Tina Dalager, Jesper Stejnicher Drongstrup Jensen, Atle Kongsvold, Anne Lovise Nordstoga, Ellen Marie Bardal, Ilya Ashikhmin, et al. Effectiveness of app-delivered, tailored self-management support for adults with lower back pain–related disability: a selfback randomized clinical trial. *JAMA internal medicine*, 181(10):1288–1296, 2021.

Nicholas Schiltz, Mary Dolansky, David Warner, Kurt Stange, Stefan Gravenstein, and Siran Koroukian. Impact of instrumental activities of daily living limitations on hospital readmission: an observational study using machine learning. *Journal of General Internal Medicine*, 07 2020. doi: 10.1007/s11606-020-05982-0.

George CM Siontis, Ioanna Tzoulaki, Peter J Castaldi, and John PA Ioannidis. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of clinical epidemiology*, 68(1):25–34, 2015.

Ewout W Steyerberg and Frank E Harrell. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology*, 69:245–247, 2016.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Modelling similarity for comparing physical activity profiles - a data-driven approach. In Michael T. Cox, Peter Funk, and Shahina Begum, editors, *CBR Research and Development*, pages 415–430, Cham, 2018. Springer. ISBN 978-3-030-01081-2.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Using automated feature selection for building case-based reasoning systems: An example from patient-reported outcome measurements. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 282–295. Springer, 2021a.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review. *Informatics*, 8(3):56, Aug 2021b. ISSN 2227-9709. doi: 10.3390/informatics8030056. URL `http://dx.doi.org/10.3390/informatics8030056`.

Deepika Verma, Kerstin Bach, and Paul Jarle Mork. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review. *Informatics*, 8(3), 2021c. ISSN 2227-9709. doi: 10.3390/informatics8030056.

Hannah J White, Jensyn Bradley, Nicholas Hadgis, Emily Wittke, Brett Piland, Brandi Tuttle, Melissa Erickson, and Maggie E Horn. Predicting patient-centered outcomes from spine surgery using risk assessment tools: a systematic review. *Current Reviews in Musculoskeletal Medicine*, 13(3):247, 2020.

Bibliography

Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, 7(1), 1994.

Amira Abdel-Aziz, Marc Strickert, and Eyke Hüllermeier. Learning solution similarity in preference-based cbr. In Luc Lamontagne and Enric Plaza, editors, *Case-Based Reasoning Research and Development*, pages 17–31, Cham, 2014. Springer International Publishing. ISBN 978-3-319-11209-1.

A Adam and H Blockeel. Dealing with overlapping clustering: A constraint-based approach to algorithm selection. *CEUR Workshop Proceedings*, 1455, 01 2015.

Daniel E Adkins. Machine learning and electronic health records: A paradigm shift. volume 174, pages 93–94. American Journal of Psychiatry, 2017. PMID: 28142275.

Abdullah Alanazi. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, page 100924, 2022.

Amani Aldahiri, Bashair Alrashed, and Walayat Hussain. Trends in using iot with machine learning in health prediction system. *Forecasting*, 3(1): 181–206, 2021. ISSN 2571-9394. doi: 10.3390/forecast3010012. URL `https://www.mdpi.com/2571-9394/3/1/12`.

Hana Alharthi. Healthcare predictive analytics: An overview with a focus on saudi arabia. *Journal of infection and public health*, 11(6):749–756, 2018.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

Gunnar BJ Andersson. Epidemiological features of chronic low-back pain. *The lancet*, 354(9178):581–585, 1999.

JA Andrews, RF Harrison, LJE Brown, LM MacLean, F Hwang, T Smith, Elizabeth A Williams, C Timon, T Adlam, H Khadra, et al. Using the nana toolkit at home to predict older adults' future depression. *Journal of affective disorders*, 213:187–190, 2017.

Muhammad Arif and Ahmed Kattan. Physical activities monitoring using wearable acceleration sensors attached to the body. *PLOS ONE*, 10(7): 1–16, 2015. doi: 10.1371/journal.pone.0130851.

Niloofar Arshadi and Igor Jurisica. Data mining for case-based reasoning in high-dimensional biological domains. *IEEE transactions on knowledge and data engineering*, 17(8):1127–1137, 2005.

Kerstin Bach and Klaus-Dieter Althoff. Developing Case-Based Reasoning Applications Using myCBR 3. In Ian Watson and Belen Diaz Agudo, editors, *ICCBR-12*. LNAI 6880, Springer, September 2012.

Kerstin Bach, Klaus-Dieter Althoff, Régis Newo, and Armin Stahl. A case-based reasoning approach for providing machine diagnosis from service reports. In *International Conference on Case-Based Reasoning*, pages 363–377. Springer, 2011.

Kerstin Bach, Klaus-Dieter Althoff, Julian Satzky, and Julian Kroehl. Cookiis mobile : A case-based reasoning recipe customizer for android phones. 2012.

Kerstin Bach, Bjørn Magnus Mathisen, and Amar Jaiswal. Demonstrating the mycbr rest api. In *ICCBR Workshops*, 2019.

Kerstin Bach, Atle Kongsvold, Hilde BÃ¥rdstu, Ellen Marie Bardal, HÃ¥kon S. KjÃŠrnli, Sverre Herland, Aleksej Logacjov, and Paul Jarle Mork. A machine learning classifier for detection of physical activity types and postures during free-living. *Journal for the Measurement of Physical Behaviour*, pages 1 – 8, 2021. doi: 10.1123/jmpb.2021-0015.

Judith Baumhauer. Patient-reported outcomes- are they living up to their potential? *New England Journal of Medicine*, 377:6–9, 07 2017. doi: 10.1056/NEJMp1702978.

Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, and Mia Folke. Case-based reasoning systems in the health sciences: a survey of recent trends and developments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(4):421–434, 2010.

Isabelle Bichindaritz and Cindy Marling. *Case-Based Reasoning in the Health Sciences: Foundations and Research Directions*, pages 127–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-14464-6.

Isabelle Bichindaritz and Stefania Montani. Introduction to the special issue on case-based reasoning in the health sciences. *Computational Intelligence*, 2009.

Isabelle Bichindaritz, Stefania Montani, and Luigi Portinale. Special issue on case-based reasoning in the health sciences. *Applied Intelligence*, 28 (3):207–209, 2008.

Asaf Bitton, Tracy Onega, Anna NA Tosteson, and Jennifer S Haas. Toward a better understanding of patient-reported outcomes in clinical practice. *The American journal of managed care*, 20(4):281, 2014.

Xiomara Blanco, Sara Rodríguez, Juan M Corchado, and Carolina Zato. Case-based reasoning applied to medical diagnosis and treatment. In *distributed computing and artificial intelligence*, pages 137–146. Springer, 2013.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Gunnar Bovim, Harold Schrader, and Trond Sand. Neck pain in the general population. *Spine*, 19(12):1307–1309, 1994.

Gunilla Brattberg, Mats Thorslund, and Anders Wikman. The prevalence of pain in a general population. the results of a postal survey in a county of sweden. *Pain*, 37(2):215–222, 1989.

Andrew M Briggs, Anthony D Woolf, Karsten Dreinhöfer, Nicole Homb, Damian G Hoy, Deborah Kopansky-Giles, Kristina Åkesson, and Lyn March. Reducing the global burden of musculoskeletal conditions. *Bulletin of the World Health Organization*, 96(5):366, 2018.

Laura JE Brown, Tim Adlam, Faustina Hwang, Hassan Khadra, Linda M Maclean, Bridey Rudd, Tom Smith, Claire Timon, Elizabeth A Williams, and Arlene J Astell. Computerized self-administered measures of mood and appetite for older adults: the novel assessment of nutrition and ageing toolkit. *Journal of Applied Gerontology*, 37(2):157–176, 2018.

J.M. Buell. The beauty of predictive analytics. leveraging data into action. *Healthcare executive*, 31:10, 2016.

Boris Campillo-Gimenez, Wassim Jouini, Sahar Bayat, and Marc Cuggia. Improving case-based reasoning systems by combining k-nearest neighbour algorithm with logistic regression in the prediction of patients' registration on the renal transplant waiting list. *PLoS ONE*, 8(9), 2013. doi: 10.1371/journal.pone.0071991.

Jean Gayton Carroll. Crossing the quality chasm: A new health system for the 21st century. *Quality Management in Healthcare*, 10(4):60–61, 2002.

Adam M Chekroud, Ralitza Gueorguieva, Harlan M Krumholz, Madhukar H Trivedi, John H Krystal, and Gregory McCarthy. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry*, 74(4):370–378, 2017.

Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon, John Harrison Krystal, and Philip Robert Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, 3(3):243–250, 2016.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Chun-Ling Chuang. Case-based reasoning support for liver disease diagnosis. *Artificial Intelligence in Medicine*, 53(1):15–23, 2011.

Patrick Clerkin, Conor Hayes, and Pádraig Cunningham. Automated case generation for recommender systems using knowledge discovery techniques. *Genre*, 1994.

Savina Colaco, Sujit Kumar, Amrita Tamang, and Vinai George Biju. A review on feature selection algorithms. In *Emerging research in computing, information, communication and applications*, pages 133–153. Springer, 2019.

Pierre Côté, J David Cassidy, Linda J Carroll, and Vicki Kristman. The annual incidence and course of neck pain in the general population: a population-based cohort study. *Pain*, 112(3):267–273, 2004.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.

Padraig Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering*, 21 (11):1532–1543, 2009a. doi: 10.1109/TKDE.2008.227.

Padraig Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Trans. Knowl. Data Eng.*, 21, 11 2009b.

S Prasanna Devi, K Suryaprakasa Rao, and S Sai Sangeetha. Prediction of surgery times and scheduling of operation theaters in optholmology department. *Journal of medical systems*, 36(2):415–430, 2012.

Harold E. Driver and Alfred Louis Kroeber. *Quantitative expression of cultural relationships, by H.E. Driver and A.L. Kroeber*. University of California Press, 1932.

Titilola O. Fanoiki, Isabela Drummond, and Sandra A. Sandri. Case-based reasoning retrieval and reuse using case resemblance hypergraphs. pages 1–7, 2010.

Mark Alan Fontana, Stephen Lyman, Gourab K Sarker, Douglas E Padgett, and Catherine H MacLean. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research*, 477(6):1267–1279, 2019.

KR Fox. At least five a week: Evidence on the impact of physical activity and its relationship to health–a report from the chief medical officer. 2004.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Al Fryan, Latefa Hamad, Mahasin Ibrahim Shomo, Malik Bader Alazzam, and Md Adnan Rahman. Processing decision tree data using internet of things (iot) and artificial intelligence technologies with special reference to medical application. *BioMed Research International*, 2022, 2022.

Thomas Gabel and Eicke Godehardt. Top-down induction of similarity measures using similarity clouds. In Eyke Hüllermeier and Mirjam Minor, editors, *Case-Based Reasoning Research and Development*, pages 149–164, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24586-7.

Aliyah Giga. How health leaders can benefit from predictive analytics. *Healthcare Management Forum*, 30:274 – 277, 2017.

Regina Guthold, Tomoko Ono, Kathleen L Strong, Somnath Chatterji, and Alfredo Morabia. Worldwide variability in physical inactivity: a 51-country survey. *American journal of preventive medicine*, 34(6):486–494, 2008.

Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15 (4):361–387, 1996.

Alex HS Harris, Alfred C Kuo, Yingjie Weng, Amber W Trickey, Thomas Bowe, and Nicholas J Giori. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research*, 477(2):452, 2019.

Jan Hartvigsen, Mark J Hancock, Alice Kongsted, Quinette Louw, Manuela L Ferreira, Stéphane Genevay, Damian Hoy, Jaro Karppinen,

Glenn Pransky, Joachim Sieper, et al. What low back pain is and why we need to pay attention. *The Lancet*, 391(10137):2356–2367, 2018.

Alec Holt, Isabelle Bichindaritz, Rainer Schmidt, and Petra Perner. Medical applications in case-based reasoning. *Knowledge Engineering Review*, 20 (3):289–292, 2005.

Eyke Hüllermeier and Patrice Schlegel. Preference-based cbr: First steps toward a methodological framework. In Ashwin Ram and Nirmalie Wiratunga, editors, *Case-Based Reasoning Research and Development*, pages 77–91, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-23291-6.

Eric L Hurwitz, Kristi Randhawa, Hainan Yu, Pierre Côté, and Scott Haldeman. The global spine care initiative: a summary of the global burden of low back and neck pain studies. *European Spine Journal*, 27(6):796–801, 2018.

Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1200–1205. Ieee, 2015.

Zhao Kang, Chong Peng, and Qiang Cheng. Kernel-driven similarity learning. *Neurocomputing*, 267:210–219, 2017.

Michael W. Kattan, Changhong Yu, Andrew J. Stephenson, Oliver Sartor, and Bertrand Tombal. Clinicians versus nomogram: predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology*, 81 5:956–61, 2013.

JENNIFER L Kelsey et al. Epidemiology and impact of low-back pain. *Spine*, 5(2):133–142, 1980.

Ronald C Kessler, Hanna M van Loo, Klaas J Wardenaar, Robert M Bossarte, Lisa A Brenner, Tianxi Cai, David Daniel Ebert, Irving Hwang, Junlong

Li, Peter de Jonge, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular psychiatry*, 21(10):1366–1371, 2016.

Kyung-Sup Kim and Ingoo Han. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. *Expert systems with applications*, 21(3):147–156, 2001.

Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992.

Harold W Kohl 3rd, Cora Lynn Craig, Estelle Victoria Lambert, Shigeru Inoue, Jasem Ramadan Alkandari, Grit Leetongin, Sonja Kahlmeier, Lancet Physical Activity Series Working Group, et al. The pandemic of physical inactivity: global action for public health. *The lancet*, 380(9838):294–305, 2012.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9): 1464–1480, 1990.

Raja Krishnamoorthi, Shubham Joshi, Hatim Z Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C Kalpana, and Basant Tiwari. A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, 2022, 2022.

I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, and Peter T Katzmarzyk. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, 380(9838):219–229, 2012. doi: 10.1016/s0140-6736(12)61031-9.

Yan Li, Simon Chi-Keung Shiu, Sankar K Pal, and James Nga-Kwok Liu. A rough set-based case-based reasoner for text categorization. *International journal of approximate reasoning*, 41(2):229–255, 2006.

Yan-Fu Li, Min Xie, and Thong Ngee Goh. A study of mutual information based feature selection for case based reasoning in software cost estimation. *Expert Syst. Appl.*, 36:5921–5931, 2009.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Jakobus M Louw, Tessa S Marcus, and Johannes FM Hugo. Patient-or person-centred practice in medicine?-a review of concepts. *African journal of primary health care & family medicine*, 9(1):1–7, 2017.

Marcos R. B. Lucca, Alcides G. Lopes Junior, Edison Pignaton de Freitas, and Luis A. L. Silva. A case-based reasoning and clustering framework for the development of intelligent agents in simulation systems. In *FLAIRS, Florida*, 2018.

Brigid M Lynch and Michael F Leitzmann. An evaluation of the evidence relating to physical inactivity, sedentary behavior, and cancer incidence and mortality. *Current Epidemiology Reports*, 4(3):221–231, 2017.

J. MacQueen. Some methods for classification and analysis of multivariate observations. Berkeley, Calif., 1967. University of California Press.

Anna Marcuzzi, Kerstin Bach, Anne Lovise Nordstoga, Gro Falkener Bertheussen, Ilya Ashikhmin, Nora Østbø Boldermo, Else-Norun Kvarner, Tom Ivar Nilsen, Gunn Hege Marchand, Solveig Osborg Ose, and et al. Individually tailored self-management app-based intervention (selfback) versus a self-management web-based intervention (e-help) or usual care in people with low back and neck pain referred to secondary care: Protocol for a multiarm randomised clinical trial. *BMJ Open*, 11(9), 2021. doi: 10.1136/bmjopen-2020-047921.

Michael Marschollek. A semi-quantitative method to denote generic physical activity phenotypes from long-term accelerometer data – the atlas index. *PLOS ONE*, 8(5), 05 2013.

Takawira C Marufu, SM White, R Griffiths, SR Moonesinghe, and Iain K Moppett. Prediction of 30-day mortality after hip fracture surgery by the nottingham hip fracture score and the surgical outcome risk tool. *Anaesthesia*, 71(5):515–521, 2016.

Bjørn Magnus Mathisen, Agnar Aamodt, Kerstin Bach, and Helge Langseth. Learning similarity measures from data. *Progress in Artificial Intelligence*, 9(2):129–143, 2020.

Suneetha Ramani Moonesinghe, Michael G Mythen, Priya Das, Kathryn M Rowan, and Michael PW Grocott. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology*, 119(4):959–981, 2013.

Giorgos Mountrakis, Peggy Agouris, and Anthony Stefanidis. Similarity learning in gis: an overview of definitions, prerequisites and challenges. *Spatial Databases: Technologies, Techniques and Trends*, pages 294–321, 2005.

Gilbert Müller and Ralph Bergmann. A cluster-based approach to improve similarity-based retrieval for process-oriented case-based reasoning. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, ECAI'14, pages 639–644, Amsterdam, The Netherlands, The Netherlands, 2014. IOS Press. ISBN 978-1-61499-418-3.

Stefano Nembrini, Inke KÃ¶nig, and Marvin Wright. The revival of the gini importance? *Bioinformatics (Oxford, England)*, 34, 05 2018. doi: 10.1093/bioinformatics/bty373.

Wendy Oude Nijeweme-d'Hollosy, Lex van Velsen, Mannes Poel, Catharina GM Groothuis-Oudshoorn, Remko Soer, and Hermie Hermens. Evaluation of three machine learning models for self-referral decision support on low back pain in primary care. *International journal of medical informatics*, 110:31–41, 2018.

Trishan Panch, Peter Szolovits, and Rifat Atun. Artificial intelligence, machine learning and health systems. *Journal of global health*, 8(2), 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2011.

H S J Picavet and A J Schuit. Physical inactivity: a risk factor for low back pain in the general population? *Journal of Epidemiology & Community Health*, 57(7):517–518, 2003. ISSN 0143-005X. doi: 10.1136/jech.57.7.517. URL `https://jech.bmj.com/content/57/7/517`.

Evan M. Polce, Kyle N. Kunze, Michael Fu, Grant E. Garrigues, Brian Forsythe, Gregory P. Nicholson, Brian J. Cole, and Nikhil N. Verma. Development of supervised machine learning algorithms for prediction of satisfaction at two years following total shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery*, 2020. ISSN 1058-2746. doi: https://doi.org/10.1016/j.jse.2020.09.007.

Sameer Quazi. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8):1–18, 2022.

Quazi Abidur Rahman, Tahir Janmohamed, Meysam Pirbaglou, Hance Clarke, Paul Ritvo, Jane M Heffernan, and Joel Katz. Defining and predicting pain volatility in users of the manage my pain app: Analysis using data mining and machine learning methods. *Journal of medical Internet research*, 20(11):e12001, 2018.

Quazi Abidur Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, and Joel Katz. Interpretability and class imbalance in prediction models for pain volatility in manage my pain app users: analysis using feature selection and majority voting methods. *JMIR medical informatics*, 7(4):e15601, 2019.

Josep Maria Raya, Daniel Montolio, and Paolo Buonanno. Housing prices and crime perception. *Empirical Economics*, 45, 08 2012. doi: 10.1007/s00181-012-0624-y.

Øyvind Reinsve and Kerstin Bach. Data analytics for hunt: Recognition of physical activity on sensor data streams, 2018.

Michael M. Richter and Rosina O. Weber. Case-based reasoning. In *Springer Berlin Heidelberg*, 2013.

Richard D. Riley, Joie Ensor, Kym I.E. Snell, Thomas P. A. Debray, Douglas G. Altman, Karel G. M. Moons, and Gary S. Collins. External validation of clinical prediction models using big datasets from e-health records or ipd meta-analysis: opportunities and challenges. *The BMJ*, 353, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Maria Salamó and Elisabet Golobardes. Rough sets reduction techniques for case-based reasoning. In *International Conference on Case-Based Reasoning*, pages 467–482. Springer, 2001.

Maria Salamo and Maite Lopez-Sanchez. Rough set based approaches to feature selection for case-based reasoning classifiers. *Pattern Recognition Letters*, 32(2):280–292, 2011.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

Louise Fleng Sandal, Mette Jensen Stochkendahl, Malene Jagd Svendsen, Karen Wood, Cecilie K Øverås, Anne Lovise Nordstoga, Morten Villumsen, Charlotte Diana Nørregaard Rasmussen, Barbara Nicholl, Kay Cooper, Per Kjaer, Frances S Mair, Gisela Sjøgaard, Tom Ivar Lund Nilsen, Jan Hartvigsen, Kerstin Bach, Paul Jarle Mork, and Karen Søgaard. An app-delivered self-management program for people with low back pain: Protocol for the selfback randomized controlled trial. *JMIR Res Protoc*, 8 (12):e14720, Dec 2019. ISSN 1929-0748. doi: 10.2196/14720.

Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah. Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)*, pages 1–6. IEEE, 2018.

Christian Severin Sauer, Lotta Rintala, and Thomas Roth-Berghofer. Knowledge formalisation for hydrometallurgical gold ore processing. In Max Bramer and Miltos Petridis, editors, *Research and Development in Intelligent Systems XXX*, pages 291–304, Cham, 2013a. Springer International Publishing.

Christian Severin Sauer, Thomas Roth-Berghofer, Nino Auricchio, and Sam Proctor. Recommending audio mixing workflows. In *ICCBR*, 2013b.

Nicholas Schiltz, Mary Dolansky, David Warner, Kurt Stange, Stefan Gravenstein, and Siran Koroukian. Impact of instrumental activities of daily living limitations on hospital readmission: an observational study using machine learning. *Journal of General Internal Medicine*, 07 2020. doi: 10.1007/s11606-020-05982-0.

Karen R. Sepucha, Steven J. Atlas, Yuchiao Chang, Andrew Freiberg, Henrik Malchau, Mahima Mangla, Harry Rubash, Leigh H. Simmons, and Thomas Cha. Informed, patient-centered decisions associated with better health outcomes in orthopedics: Prospective cohort study. *Medical Decision Making*, 38(8):1018–1026, 2018. doi: 10.1177/0272989X18801308.

Zahra Shahabi Kargar, Sankalp Khanna, Norm Good, Abdul Sattar, James Lind, and John O'Dwyer. Predicting procedure duration to improve scheduling of elective surgery. In *Pacific Rim International Conference on Artificial Intelligence*, pages 998–1009. Springer, 2014.

Hon-Yi Shi, Jinn-Tsong Tsai, Yao-Mei Chen, Richard Culbertson, Hong-Tai Chang, and Ming-Feng Hou. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. *Breast cancer research and treatment*, 135(1):221–229, 2012.

Patricia L Sinnott, Sharon K Dally, Jodie Trafton, Joseph L Goulet, and Todd H Wagner. Trends in diagnosis of painful neck and back conditions, 2002 to 2011. *Medicine*, 96(20), 2017.

Barry Smyth. Recommender systems: A healthy obsession. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9790–9794, 2019.

Barry Smyth and Pádraig Cunningham. Running with cases: A cbr approach to running your best marathon. In David W. Aha and Jean Lieber, editors, *Case-Based Reasoning Research and Development*, pages 360–374, Cham, 2017. Springer International Publishing. ISBN 978-3-319-61030-6.

Armin Stahl. Learning feature weights from case order feedback. In *IN PROCEEDINGS OF THE 4TH INTERNATIONAL CONFERENCE ON CASE-BASED REASONING*, pages 502–516. Springer, 2001.

Armin Stahl. Learning similarity measures: A formal view based on a generalized cbr model. In *OPTIONAL COMMENT/QUALIFICATION: IST-1999-10357/BRI/WP5/0230 © FORM CONSORTIUM D10: VALIDATION OF INTER-ENTERPRISE MANAGEMENT FRAMEWORK (TRIAL 2) – ANNEX B PAGE 24 OF 29 12. HOW*, pages 507–521. Springer, 2005.

Armin Stahl and Thomas Gabel. Using evolution programs to learn local similarity measures. In *IN PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON CASE-BASED REASONING*, pages 537–551. Springer, 2003.

Armin Stahl and Thomas Gabel. Optimizing similarity assessment in case-based reasoning. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 21, page 1667. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Armin Stahl and Thomas R. Roth-Berghofer. Rapid prototyping of cbr applications with the open source tool mycbr. In *ECCBR '08*. Springer-Verlag, 2008. ISBN 978-3-540-85501-9.

Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

Mobyen Uddin and Amy Loutfi. Physical activity identification using supervised machine learning and based on pulse rate. *International Journal of Advanced Computer Science and Applications*, 4(7), 2013. doi: 10.14569/ijacsa.2013.040730.

Martin B Van Der Weyden. Databases and evidence-based medicine in general practice. *Medical journal of Australia*, 170(2):52–53, 1999.

Gabriel R Vasquez-Morales, Sergio M Martinez-Monterrubio, Pablo Moreno-Ger, and Juan A Recio-Garcia. Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning. *Ieee Access*, 7:152900–152910, 2019.

Filip Velickovski, Luigi Ceccaroni, Josep Roca, Felip Burgos, Juan B Galdiz, Nuria Marina, and Magí Lluch-Ariet. Clinical decision support systems (cdss) for preventive management of copd patients. *Journal of translational medicine*, 12(2):1–10, 2014.

Eirik Vågeskar. Activity recognition for stroke patients, Jul 2017. URL `https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2468160`.

Xin Shelley Wang and Vijaya Gottumukkala. Patient-reported outcomes: Is this the missing link in patient-centered perioperative care? *Best Practice & Research Clinical Anaesthesiology*, 35(4):565–573, 2021.

Yan Wang, Ronald D Hays, Marvin Marcus, CA Maida, J Shen, D Xiong, ID Coulter, SY Lee, VW Spolsky, JJ Crall, et al. Developing children's oral health assessment toolkits using machine learning algorithm. *JDR Clinical & Translational Research*, 5(3):233–243, 2020.

Rosina Weber, Jason M. Proctor, Ilya Waldstein, and Andres Kriete. Cbr for modeling complex systems. volume 3620, 08 2005.

Nirmalie Wiratunga, Susan Craw, and Stewart Massie. Index driven selective sampling for cbr. In Kevin D. Ashley and Derek G. Bridge, editors, *Case-Based Reasoning Research and Development*, pages 637–651. Springer Berlin Heidelberg, 2003.

DJN Wong, CM Oliver, and SR Moonesinghe. Predicting postoperative morbidity in adult elective surgical patients using the surgical outcome risk tool (sort). *BJA: British Journal of Anaesthesia*, 119(1):95–105, 2017.

Ning Xiong and Peter Funk. Construction of fuzzy knowledge bases incorporating feature selection. *Soft Computing*, 10(9):796–804, 2006.

Ning Xiong and Peter Funk. Combined feature selection and similarity modelling in case-based reasoning using hierarchical memetic algorithm. In *IEEE Congress on Evolutionary Computation*, pages 1–6. IEEE, 2010.

Zhirong Yang, Jukka Cor, er, and Erkki Oja. Low-rank doubly stochastic matrix decomposition for cluster analysis. *Journal of Machine Learning Research*, 17(187), 2016.

Bangpeng Yao and Shao Li. Anmm4cbr: a case-based reasoning method for gene expression data classification. *Algorithms for Molecular Biology*, 5 (1):14, 2010. doi: 10.1186/1748-7188-5-14.

Guo-Niu Zhu, Jie Hu, Jin Qi, Jin Ma, and Ying-Hong Peng. An integrated feature selection and cluster analysis techniques for case-based reasoning. *Engineering Applications of Artificial Intelligence*, 39:14–22, 2015.

Zoe Y Zhuang, Leonid Churilov, Frada Burstein, and Ken Sikaris. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. *European Journal of Operational Research*, 195(3):662–675, 2009.

Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The feature importance ranking measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 694–709. Springer, 2009.

ISBN 978-82-326-6543-3 (printed ver.)
ISBN 978-82-326-5426-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology