Strategies for Generalisable Machine Learning with Small Data for Exercise Fatigue Detection

by

André Jeworutzki

Thesis submitted in partial fulfilment of the requirements of the University of the West of Scotland for the award of Doctor of Philosophy

January 1, 2025

Declaration

The research presented in this thesis was carried out by the undersigned. No part of the research has been submitted in support of an application for another degree or qualification at this or another university.

January 1, 2025

Heroes

Kai von Luck, Jan Schwarzer, Susanne Draheim, Tobias Eichler, Jessica Broscheit, Qi Wang, Lukas Weiler, Mike Schmidt, Marcelle Schaffarczyk, Max Kolwa, and Chris Easton

Fatigue is the most comfortable pillow.

Abstract

This thesis contributes to the field of human-centred computing by exploring strategies and pitfalls for developing generalisable machine learning models for sensor-based exercise fatigue detection with small data. Machine learning faces several challenges in exercise fatigue detection due to the scarcity of available data, as fatigue research typically relies on sample data from a limited number of subjects due to the time and effort required to design, conduct, and analyse studies with human subjects. Although machine learning models can provide robust predictions, even when trained on small data sets, careful consideration of variability and data distribution is required to improve their generalisability.

A step-by-step framework for exercise fatigue detection with machine learning and small data is introduced in this thesis. The framework is implemented in a case study of 48 subjects performing squat exercises, using inertial measurement units and 2D pose estimation to capture movement patterns and correlate these patterns with ratings of perceived exertion. The results are analysed in terms of generalisability, including different numbers of classes and subjects, class imbalances, k-fold cross-validation, oversampling, inter-individual variability, evaluation metrics, and evaluation types.

Based on a comparison between inertial units and 2D pose estimation, it is concluded that 2D pose estimation can be used for fatigue detection. From the literature survey and case study, it is also concluded that most exercise fatigue detection models trained on small data sets may not perform well in a real-world application.

Abbreviations

A list of the key abbreviations used throughout this thesis:

HAR Human Activity Recognition

Ability to detect human activity or condition based on information from sensors.

PE Pose Estimation

A computer vision task that predicts and extracts the location of key points (joints) for one or more individuals. PE-Side and PE-Front refer to PE based on the side and front camera perspective, respectively.

IMU Inertial Measurement Unit

An IMU is an electronic sensor device that incorporates a combination of accelerometer, gyroscope, and sometimes magnetometer.

RPE Rating of Perceived Exertion

RPE is based on a qualitative scale by Borg [51], commonly used in sport research.

ML Machine Learning

ML gives computers the ability to learn without being explicitly programmed. It draws on concepts from several scientific disciplines, including linear algebra, optimisation problems, probability theory, statistics, and artificial intelligence.

T1-SOLO Evaluation Type 1: Personalised, Single Subject

Data from one individual is used for ML.

T2-LNSO Evaluation Type 2: Leave No Subject Out

Some data from all subjects is used as test set and excluded from the training set.

T3-LOSO Evaluation Type 3: Leave One Subject Out

Data from one subject is used as test set and excluded from the training set.

T4-LMSO Evaluation Type 4: Leave Multiple Subjects Out

Data from multiple subjects is used as test set and excluded from the training set.

Contents

List of Figures

List of Tables

1 Introduction

During the COVID-19 pandemic, attendance at fitness studios and rehabilitation centres was restricted, making unsupervised home exercise the only feasible option for many people [334]. As a result, people have become accustomed to other forms of exercise training, such as remote or remotely monitored home exercise [93, 16, 127]. However, home exercise proves challenging in regard to feedback on exercise performance, which is an important area of research in sport and healthcare [239, 297]. A monitoring system can act as an early warning mechanism [239, 144] to prevent injury, monitor the effectiveness of the training programme, maintain performance, and prevent overtraining [239, 143, 323].

In computer science, *human activity recognition* (HAR) is the ability to recognise human activity based on information from various sensors, which may include cameras, wearable sensors, sensors attached to everyday objects, or sensors placed in the environment [162]. HAR has the potential to monitor and support exercise training at home, particularly for personal fitness and rehabilitation, in the absence of a personal trainer [66, 40, 147]. In this context, *machine learning* (ML) has attracted the attention of researchers in healthcare, sports science, and HAR to improve the performance of assistive exercise training systems [343, 368]. For example, ML methods may complement established models, such as the Fitness-Fatigue model used in sports science [61], by incorporating more complex physiological representations and using non-linear, multivariate algorithms [163].

Fatigue detection during physical activity can be considered as a special case of HAR [98]. Exercise training is associated with physical fatigue [247], which refers to the sensation of physical exhaustion resulting from physical exertion [276, 15]. Its presence not only increases the risk of injury but also reduces exercise performance [256, 240, 306, 247, 189, 108]. Early detection of fatigue during

physical activity can prevent overexertion, illness, and injury; and help individuals adjust their activities or schedules accordingly [217, 306, 175, 143].

However, ML faces several challenges in the field of fatigue detection in physical exercise, mainly due to the scarcity or inadequacy of data [100]. Although ML models typically perform best when trained on large amounts of data [369, 354], fatigue research often relies on sample data from a small number of subjects, hereafter referred to as small data [150]. Most ML approaches to HAR rely on supervised learning based on labelled data [62]. The main reason for small data is the time and effort required to design and conduct studies with human subjects, followed by the effort to clean, label, and analyse the collected data; this process typically involves a large amount of manual labour and semi-automated methods [62, 100, 204, 162]. ML models can still provide robust predictions even when trained on small data sets, but careful consideration of variability and data distribution is necessary to improve generalisability [100]. There are techniques for dealing with small data, such as transfer learning, regularisation, and visualisation, but they require skilled practitioners and their effectiveness can be limited [204].

This thesis is divided into three main parts. The first part is about a literature review to identify common strategies and pitfalls in sensor-based exercise fatigue detection with ML and small data in order to design a Fatigue Recognition Chain framework. This framework is intended as a general guide for interdisciplinary researchers to conduct similar research projects. The framework covers the entire process of building an exercise fatigue detection system from specification, collecting raw data, data transformation, ML, evaluation, and dissemination of the results. The second part is a case study, conducted as part of this thesis, with 48 subjects to demonstrate the implementation of the framework and to collect data during squat exercise based on *ratings of perceived exertion* (RPE), *Inertial Measurement Unit* (IMU), and *pose estimation* (PE). The collected data is processed to train different ML models, which are analysed for their generalisability under different evaluation

methods, imbalanced data sets, and augmentation techniques. The third part is a synthesis and discussion of the results of the two previous parts.

1.1 Research Aim

The aim of this thesis is to identify and address the strategies and pitfalls of sensor-based exercise fatigue detection using ML with small data sets for physical activities such as exercise training in terms of generalisability.

1.1.1 Research Questions

To fulfil this aim, the following research questions are investigated:

1. How to conduct research on exercise fatigue detection with ML?
2. What are common strategies and pitfalls of ML with small data?
3. How do small data, evaluation methods, and augmentation effect ML?
4. How generalisable are ML models trained on small data sets?

1.1.2 Research Objectives

To accomplish the research questions, the following objectives are defined:

1. To review the literature on exercise fatigue detection based on sensors and ML.
2. To create a framework for sensor-based fatigue detection research with ML.
3. To conduct a case study with squat exercises by implementing the framework.
4. To collect RPE-labelled sensor data from IMU and PE for ML analyses.
5. To investigate the ML fatigue predictions with an increasing data set.
6. To compare evaluation types and their effect on generalisability.
7. To explore data augmentation techniques to improve generalisability.

1.2 Research Contributions

This thesis contributes knowledge to the field of human-centred computing, which studies the design, development, and deployment of mixed-initiative human–computer systems and has emerged from the convergence of several disciplines concerned with the understanding of humans and the design of computational artefacts [167]. The contributions are supported by the following peer-reviewed publications:

1. "A Preliminary Experimental Outline to Train Machine Learning Models for the Unobtrusive, Real-Time Detection of Acute Physiological Stress Levels during Training Exercises" [173]
2. "Determining acute physiological stress levels with wearable sensors based on movement quality and exhaustion during repetitive training exercises" [172]
3. "Small Data, Big Challenges: Pitfalls and Strategies for Machine Learning in Fatigue Detection" [174] – Winner of the "Best Student Paper Award"

Regarding the research objectives, publication 1 covers the results of the first literature review and case study, as well as a first version of a framework. Publication 2 focused on the analysis of the ML models based on the collected data. Publication 3 used the collected data from additional case studies to analyse the fatigue predictions with an increasing data set and to compare different evaluation types and augmentation techniques for their generalisability.

Table A.1 in the Appendix illustrates the contributions of this thesis in relation to the related works. These are briefly described below:

Survey and Analysis of Generalisability and Evaluation Methods Some literature reviews addressed the use of ML to detect fatigue [157] or exercise fatigue [247, 248], but did not examine generalisability or evaluation methods. While many primary studies have used ML with small data to detect exercise fatigue, discussions of the generalisability are limited. The lack of strategies for developing and evaluating generalise ML models with small data highlights a substantial gap in current research (see Table A.1 in Appendix A). Addressing this gap is critical to

avoid common pitfalls and to train more reliable and applicable ML models for fatigue detection.

Framework for Exercise Fatigue Detection with ML and Small Data General approaches exist for HAR [57] and fatigue detection [50]. Maman et al. [242] proposed a framework for physical fatigue management using wearable sensors and ML. However, there is no comprehensive guidance that specifically addresses the intersection of HAR and fatigue detection with respect to ML and small data. This gap highlights the need for a focused investigation of methods and strategies tailored to these combined domains (see Table A.1). Future studies can draw on this knowledge to design their own research and contribute to their field. Practitioners, such as gym staff and health equipment developers, may benefit from the development of unobtrusive fatigue detection systems. This research could also lead to less expensive equipment and less labour-intensive ways of detecting fatigue during exercise, for example, by avoiding the need for blood samples. Improved and timely feedback to exercisers can enrich their training experience and help prevent injuries. Feedback on the quality of exercise performance is vital in sports and healthcare [297].

PE for Exercise Fatigue Detection To date, no study has utilised PE based on 2D cameras, nor has any study compared the effectiveness of ML models using data from IMUs and PE for the purpose of exercise fatigue detection (see Table A.1). This research aims to fill this gap by evaluating the performance of these data sources and providing insights into their relative strengths and weaknesses in detecting fatigue during exercise.

1.3 The Researcher's Role

Creswell [88] emphasised the need to clarify the role of the researcher to ensure the credibility of the research. This section describes the author's background and motivation for this work. This thesis started as part of the interdisciplinary research

project “MoGaSens”[1], which was funded by the European Regional Development Fund through the Hamburgische Investitions- und Förderbank from 2019 to 2022. The aim of the project was to develop a smart training shirt for the home fitness market, capable of assessing movements during training exercises in real time and providing additional health information. The experiments were carried out in the Creative Space for Technical Innovations[2] at the Hamburg University of Applied Sciences[3], Faculty of Engineering and Computer Science, where the author was employed on a full/part-time basis.

The author is a yoga teacher with an additional focus on therapeutic yoga, aimed at preventing injuries during yoga practice. The author’s interest in MoGaSens was driven by the project’s goal of preventing injuries during exercise through computational support. During the research project, the challenge of determining the amount of data needed to train ML models was a recurring theme. As a computer scientist, this became a key question, as machine support for injury prevention relies on the generalisation and reliability of these models.

Since 2021, the author has also been teaching the course "Train Like A Machine"[4] every semester, where groups of students explore different sensors and ML methods for analysing physical activity. These courses have so far resulted in two publications [329] and [155]. In addition, the author has co-supervised three bachelor theses in this context.

1.4 Research Timeline

The following timeline, shown in Figure 1.1, presents an overview of the key milestones, research activities, and outputs achieved during the course of this thesis. The research outputs are marked by three key publications. The timeline

[1] https://csti.haw-hamburg.de/project/mogasens/
[2] https://csti.haw-hamburg.de
[3] https://haw-hamburg.de
[4] https://csti.haw-hamburg.de/project/tlam/

illustrates the temporal alignment of the research activities, offering a narrative of the development and trajectory of the thesis.

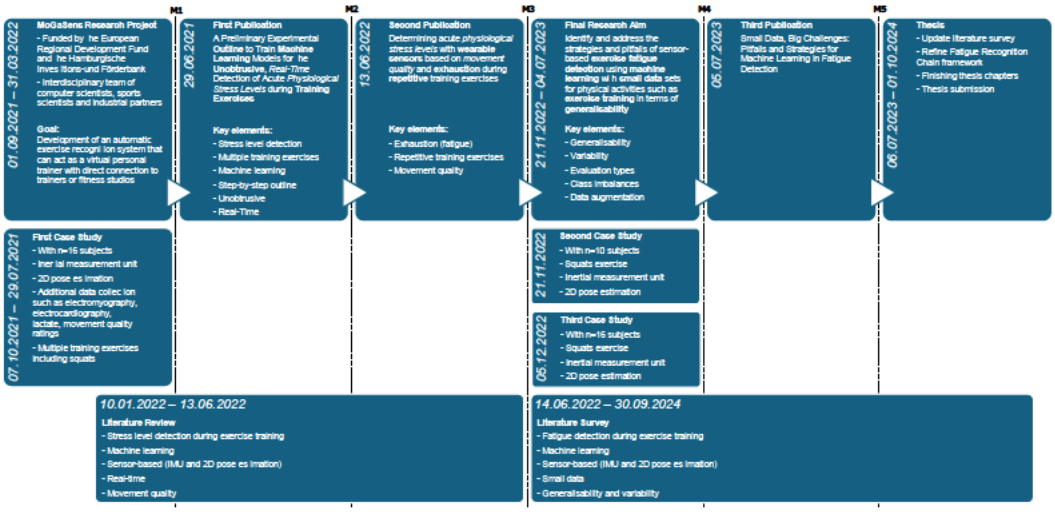


Fig. 1.1.: Research timeline of this thesis.

An interdisciplinary research project called MoGaSens was the basis for this research. The research project was carried out between September 2021 and March 2022. It aimed to develop an automatic training recognition system that can act as a virtual personal trainer. The research includes the first case study which was conducted between October 2021 and July 2022, involved 16 participants and used multi-modal data collection techniques such as inertial measurement units, 2D pose estimation, and physiological metrics such as lactate, electrocardiography, and electromyography (milestone M1).

Literature reviews were a critical part of the research process. From January to June 2022, the research focused on sensor-based methods for detecting stress levels during exercise, where stress is a more abstract phenomenon that includes fatigue.

Based on the literature review and the results of the research project, two peer-reviewed publications were published. The first focused on a methodological framework for exercise stress detection with ML (milestone M2) and the second focused on the detection of multiple stress levels during exercise training (milestone M3).

To investigate the generalisability of the trained machine learning models from the first case study, a second case study was conducted in December 2022, focusing on squat exercises. Shortly after, a third case study revisited the squat exercises to further validate the findings. Milestone M3 represents a turning point, as the analysis of the data collected led to a shift in focus to the challenges of fatigue detection, including small data, generalisability, and variability. The shift from stress to fatigue detection was a measure to narrow the research topic (milestone M4).

The final phase of the research, from November 2022 to October 2024, focused on identifying strategies and addressing pitfalls in sensor-based fatigue detection using small data. This phase focused on overcoming issues such as generalisability, class imbalances, and small data evaluation. The findings were published in a third peer-reviewed publication (milestone M5).

Finally, the thesis was completed by updating the literature survey, refining the Fatigue Recognition Chain framework, finalising the chapters, and preparing for submission.

1.5 Key Terms

This section provides definitions for key terms used throughout this thesis.

Fatigue Fatigue is the momentary sensation of feeling the need for physical rest or the mismatch between expended physical effort and actual physical performance.

Data points Data points are individual units of raw data (signals).

Small data Small data refers to data points from a small number of different individuals.

Time series A time series is a sequence of data points collected at regular intervals over time. Time series data allows the analysis of changes in a variable over a period of time. Timestamps are often included to represent the order of the sequence, but they do not necessarily have to represent the actual time [166].

Features Features (or attributes) are low-level properties of data points that can be measured or automatically calculated [179].

Samples Samples are individual instances or observations; each instance is typically represented as a vector (row) of multiple features, with each column corresponding to a specific feature (dimension).

ML Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed [291]. It draws on concepts from several scientific disciplines, including linear algebra, optimisation problems, probability theory, statistics, and artificial intelligence [179].

ML model An ML model is a mathematical construct derived from data points or features that is designed to recognise patterns and make automated predictions.

ML method An ML method is an algorithm to train ML models.

Individual-based ML Individual-based ML aims to predict the behaviour or state of individuals (or subjects) rather than treating the entire data set as a homogeneous group, recognising that each individual has unique characteristics and patterns.

Generalisation Generalisation in ML is the ability of a trained ML model to accurately predict results on unseen data which is assessed by testing [363].

Overfitting Overfitting ML models have low training but high testing error [148].

Label / Class / Target variable Label, class, and target variable are terms used interchangeably in ML and refer to the output variable that the ML model is trying to learn to predict.

Class distribution Class distribution is the proportion of samples belonging to each class (label) in a data set.

RPE Rating of Perceived Exertion is based on a qualitative scale developed by Borg [51] and commonly used in sport science.

1.6 Thesis Structure

The following chapters structure this thesis:

Chapter 1: Introduction This chapter provides the contextual background for this thesis, outlining the research aim, questions, and objectives. It also presents the research contributions, key terms used throughout this thesis, and the researcher's role, concluding with this overview.

Chapter 2: Literature Review This chapter reviews the existing literature on the key concepts in this thesis: small data, fatigue detection and HAR. It covers definitions, taxonomies and measurement techniques. Following the key concepts, primary studies and research gaps are presented.

Chapter 3: Fatigue Recognition Chain Framework This chapter introduces the Fatigue Recognition Chain framework which draws on the works identified in the literature review. General approaches and challenges are discussed in seven steps, including the definition of foundational characteristics, raw data collection, preprocessing, data transformation, feature engineering, ML, evaluation, and dissemination of research results.

Chapter 4: Case Study: Fatigue Detection for Squats with IMU and PE This chapter describes the case study conducted for this thesis, including a description of the research design, research setting, sample selection, and data collection procedures. It demonstrates the implementation of the fatigue recognition chain framework specifically for squat exercises.

Chapter 5: Results This chapter presents the results of the experiments from the case study. It discusses the performance of the trained ML models, the choice of data source (IMU and PE) for fatigue detection, the role of RPE thresholds and class number, feature augmentation for imbalanced classes, and the effect of evaluation types. In addition to classification, regression methods are investigated. Furthermore,

the performance of the ML models with increasing number of subjects is analysed to investigate their generalisation.

Chapter 6: Discussion This chapter discusses the results, including a comparison between IMU and PE data, the strategies and pitfalls of generalisable ML for exercise fatigue detection, and generalisability myths. It also summarises the findings and implications of this thesis.

Chapter 7: Conclusion The final chapter summarises the research, including its implications and limitations. Recommendations for future research directions are also provided.

"In 1976, George Box, a famous statistician, said: all models are wrong but some are useful. He was talking about statistical models, but the same is true today for machine learning (ML) models." – Baeza-Yates [25]

2 Literature Review

This chapter is divided into two parts. The first part presents a comprehensive literature review of the core concepts of this thesis: small data, fatigue detection, and HAR. Generalisation is addressed in later chapters.

The second part is a systematic review of primary studies on sensor-based fatigue detection with ML and physical activities. This part provides statistics on the identified studies, including ground truth, imbalanced data, ML methods, evaluation techniques, augmentation techniques as well as how small data and generalisation are addressed. Based on this analysis, research gaps are highlighted.

The findings of this literature review form the basis for the development of the Fatigue Recognition Chain framework presented in the following Chapter 3.

2.1 Small Data

While the term small data has been noted in publications as early as 1989, the majority of the earlier publications used the term to refer to small data sets, typically from the field of statistical mathematics, as opposed to the more recent variations defined in the field of data and information sciences [324]. The first two publications on the use of ML on small data sets indexed in Scopus were published in 1995. After that, publications were rare until 2002, with the trend starting to increase linearly in 2003 and exponentially in 2016 [202].

2.1.1 Small Data Definition(s)

In the context of this thesis, small data refer to data sets collected from a small number of different individuals that are used for ML to make predictions at the individual level [150].

However, several other definitions for small data exist as well as perspectives [198, 324, 115]. For example, Thinyane [324] explored the concept of small data, including various interpretations and perspectives on what constitutes small data:

- Small data sets: This perspective emphasises the size of the data sets used, contrasting them with the characteristics of big data.
- Actionable by-products of big data analytics: This perspective focuses on the valuable insights generated by big data analytics.
- n = me: This perspective emphasises the individual-centric nature of small data, often associated with digital traces left by individuals.
- Ethnographic human-centric observations: This approach emphasises the importance of gaining insights into human behaviour and preferences.
- An approach to data analysis: The unit of analysis of data is congruent to the unit of sampling of the data (e.g., individual-, household-, or city-level).

Miller [251] pointed out that small data studies often rely on tightly controlled sampling techniques. These techniques can limit the scope, temporality, size and diversity of the data, as well as the ability to capture and define levels of error, bias, uncertainty, and provenance. Small data are therefore characterised by their limited volume, non-continuous collection, limited diversity, and are usually generated to answer specific questions.

Rauschenberger and Baeza-Yates [284] stated that small data in data science might refer to 15000 data points for image analysis, whereas in human-centred design it might refer to around 200 or less participants, depending on the domain and context.

2.1.2 Small vs Big Data

Another approach to defining small data is to compare small data to big data. According to Kitchin and Lauriault [198], the term big is misleading because big data is characterised by more than volume. Some small data sets can be very large, such as the national censuses. However, census data sets lack velocity (typically conducted once every 10 years), variety (typically about 30 structured questions), and flexibility (it is almost impossible to change the questions). Other small data sets also consist of a limited combination of the characteristics of large data sets. For example, a qualitative data sets, such as interview transcripts, tend to be relatively small in size, non-continuous in temporality, weakly relational, and limited in variety, but high in resolution and flexibility. A comparison of different characteristics of small and big data is shown in Table 2.1.

Tab. 2.1.: Comparison of small and big data characteristics by Kitchin and Lauriault [198].

Characteristic	**Small data**	**Big data**
Volume	Limited to large	Very large
Exhaustivity	Samples	Entire populations
Resolution and indexicality	Coarse and weak to tight and strong	Tight and strong
Relationality	Weak to strong	Strong
Velocity	Slow, freeze-framed	Fast
Variety	Limited to wide	Wide
Flexible and scalable	Low to middling	High

Kong et al. [204] noted that in the natural sciences, annotated data sets tend to be small because data is typically collected manually using sophisticated equipment. Despite the growing interest in big data in recent years, many problems are small data problems [25, 202, 148, 115]. Kitchin and Lauriault [198] criticised the focus on big data, which does not make the scientific method obsolete: data cannot be analysed without hypotheses. Although ML finds patterns where science cannot, correlation does not replace causation, and science should not proceed without coherent models, unified theories, or any explanation at all.

Faraway and Augustin [115] pointed out that asymptotic analysis, while theoretically valuable, can have practical limitations with large data sets. Confidence

intervals can become too narrow, leading to overly certain conclusions. Bayesian approaches can also face this problem, as the likelihood can dominate the prior. ML practitioners often avoid dealing with uncertainty by not providing explicit estimates of uncertainty. However, uncertainty arises from factors beyond unknown parameters, such as model selection and data quality issues. Incorporating these uncertainties into models can improve their realism, but it's a challenging task.

2.1.3 Small Data Quality

According to Lauriault [222], due to the limited sample sizes of small data, data quality is paramount. They defined the following characteristics: small data sets are clean (free of errors and gaps), objective (unbiased and representative of the real world), consistent (with minimal discrepancies or inconsistencies), veracious (authentic and accurately representing what it is intended to represent), and well-documented (with clear lineage and provenance to establish its suitability for use).

2.1.4 Small Data Benefits

Small data sets have a long history of development, with established methodologies and analysis techniques. Small data can be tailored to specific research questions, allowing in-depth exploration of individual interactions and the complex ways in which people make sense of the world. Researchers can focus on specific cases, providing detailed, nuanced and contextualised stories. Small data can provide valuable insights that may be missed by big data analysis methods [198]. According to Faraway and Augustin [115], researchers often prefer small data sets, collected under controlled experimental conditions, to large observational data of unknown origin – where an inference of causality is desired, quality of data beats quantity.

2.1.5 Small Data Challenges

According to Kokol et al. [202], the most commonly reported challenges related to small data are small data set size, high/low dimensionality, and imbalanced data. Small data set size is common due to the high cost of sampling [202, 166, 162, 154, 260] and imbalanced data is a long-standing problem in ML [373, 204]. Moreover, aggregation bias can occur when large individual data sets are reduced to smaller groups, which is particularly relevant in fields like personalised medicine [115].

ML often performs poorly on small data sets [115, 231]. ML, especially deep learning, can learn effectively on big data sets, but cannot learn effectively on small data sets due to problems such as overfitting, noise, outliers, and sampling bias [204, 154]. Kong et al. [204] highlighted overfitting as a major problem in the analysis of small data, where a solution has to be found from a relatively large hypothesis space with insufficient heuristics (guidance) in the form of data. The ability of ML to detect patterns is proportional to the size of the data set; the smaller the data set, the less powerful and accurate ML methods are. Kong et al. [204] also made the distinction between data and knowledge. When there is big data, ML needs a small amount of knowledge. When there is small data, ML needs a large amount of knowledge to reduce the model search space. How to extract and represent knowledge to support ML is a major challenge.

A data set has to be representative of the cases to be predicted. Complex ML models can detect subtle patterns in the data, but noise or small data sets can mislead models into detecting patterns in the noise itself [128]. Communicating uncertainty is crucial with small data sets, as certain regions may be under-represented, affecting the model's ability to generalise [154]. However, finding a data set that comprehensively covers all possible matches is almost impossible [62].

2.1.6 Small Data Strategies

There are strategies for dealing with small data, such as data augmentation, transfer learning, regularisation and visualisation. However, these methods require skilled practitioners and their effectiveness may be limited [204, 154]. Other strategies for dealing with small data include reducing the number of possible hypotheses, reducing the degrees of freedom, and reducing the complexity of the models [154]. Another strategy may be to develop custom ML methods specialised for small data [154], for example, model-based ML [204], an alternative Naive Bayes approach [85], context trees [105], multiple runs of neural networks, customised decision trees [303], or adaptive local hyperplane algorithms [356].

2.2 Fatigue Detection

Fatigue is a multifaceted phenomenon that has been studied in various research fields[1], such as cognitive neuroscience, exercise physiology, psychology, medicine, and workplace fatigue [217, 311, 269]. This section explores the different definitions, factors, classifications, and measures of fatigue, as well as the ongoing challenges in understanding and assessing fatigue, including both subjective perceptions and objective changes in activity performance.

2.2.1 Fatigue Definition(s)

There are many definitions of fatigue [217, 43, 248] and there are ongoing attempts to unify existing definitions [217, 311, 269, 110, 199]. One of the main obstacles has been the scope of its usage: fatigue can denote a reduction in physical and cognitive function, ranging from exercise-induced impairment of motor performance to feelings of tiredness and weakness that may be accompanied by clinical conditions [110]. According to Enoka and Duchateau [110], it is not possible to identify the

[1] A discussion of how fatigue relates to stress can be found in the Appendix E.

etiology (causes or origins) of fatigue by attempting to disentangle the decline in muscle force from sensations about fatigue. Exertion and fatigue are states with both physiological and psychological aspects [114, 50]. Kluger et al. [199] argued that when dealing with fatigue, one should distinguish between *subjective perceptions* (fatigue) and *objective changes in performance* (fatigability). Pattyn et al. [269] proposed three essential components of fatigue: the perception of effort, the propensity to exercise effort which is the product of a decision-making process, and the motivation which depends on several factors and influences the propensity to exercise effort.

Fatigue is often classified into different types: physical, mental, cognitive, or emotional [311]. Billones et al. [43] also identified motivational, central, peripheral, and psychosocial fatigue. They found that 83% of the reviewed studies assessed multiple types of fatigue at the same time. For example, Elsais et al. [107] characterised physical fatigue – which is the focus of this thesis – by muscle fatigability: the difficulty to initiate or sustain muscle activities. Exercise-induced fatigue (i.e., the inability to continue a given exercise) is often associated with peripheral and central factors [269]. Peripheral fatigue is usually described as an impairment located in the muscle and characterised by a metabolic end point, while central fatigue is defined as a failure of the central nervous system to adequately drive the muscle [269].

Martins et al. [248] categorised fatigue into the following four types: mental fatigue, drowsiness, physical fatigue, and muscle fatigue. Mental fatigue is the decrease in mental performance as a result of cognitive overload (due to task duration and/or workload), independent of sleepiness. Drowsiness is fatigue arising from sleep- and circadian rhythm-related factors (e.g., sleep deprivation, circadian rhythm disruption), monotony or low task workload. Physical fatigue is the decline in overall physical performance caused by physical exertion. Muscle fatigue is the decrease in an isolated muscle performance due to reduced contractile activity.

Fatigue can also be distinguished by the time frame in which it occurs. State fatigue is the momentary (acute) sensation of fatigue and can change rapidly within

minutes or hours, whereas trait fatigue is the overall disposition and intensity of fatigue over a period of time, i.e., each individual always has trait fatigue to a varying degree [311, 110]. Prolonged state fatigue also involves the effect of recovery [311]. Chronic fatigue is generally defined as fatigue above a certain level that lasts for six months or more [107]. Pathological fatigue is based on an identifiable cause, consequence, or result of a disease, disorder, or trauma, e.g., cancer-related fatigue [311]. Furthermore, some researchers distinguish between active and passive fatigue. Passive fatigue is caused by prolonged, monotonous, boring work, whereas active fatigue is caused by prolonged task-related work [269].

According to Enoka and Duchateau [110], fatigue is a single entity that does not need to be modified by accompanying adjectives such as central fatigue, mental fatigue, or muscle Although such descriptors are usually intended to imply the likely location of the modulating factors that limit performance, the distinctions are too vague to be meaningful and lead to an incoherent literature on fatigue. The following composite definition of fatigue is used in this thesis. It is based on the set of 13 definitions identified by Skau et al. [311]:

> Fatigue is the momentary sensation of feeling the need for physical rest or the mismatch between expended physical effort and actual physical performance.

Where physical rest is a beneficial state that is intentional, temporary, and restorative, involving cessation, minimisation, or change in activity or well-being (modified definition based on Bernhofer [36]). Physical effort is the expenditure of energy for the purpose of setting the body in motion (definition derived from Massin [249]). Physical performance is any bodily activity that can be rated (e.g., by a jury) or measured (e.g., by time, length, weight, or counting).

With regard to physical effort (e.g., exercise), fatigue is an inevitable consequence [108]. Exercise-induced fatigue occurs when the effort required by the exercise task equals the maximum effort that the individual is willing to exert to succeed in the

task, or when the individual believes to have exerted a true maximum effort and continuation of the exercise is perceived as impossible [269].

Based on this definition, fatigue can be captured by movement patterns using sensors such as IMU and pose estimation to correlate these patterns with subjective ratings of perceived exertion.

2.2.2 Fatigue Taxonomies

Enoka and Duchateau [110] proposed a taxonomy of different factors influencing fatigue derived from two interdependent attributes: perceived fatigability and performance fatigability. The former consists of homeostasis and psychological state, the latter of contractile function and muscle activation (see Figure 2.1).

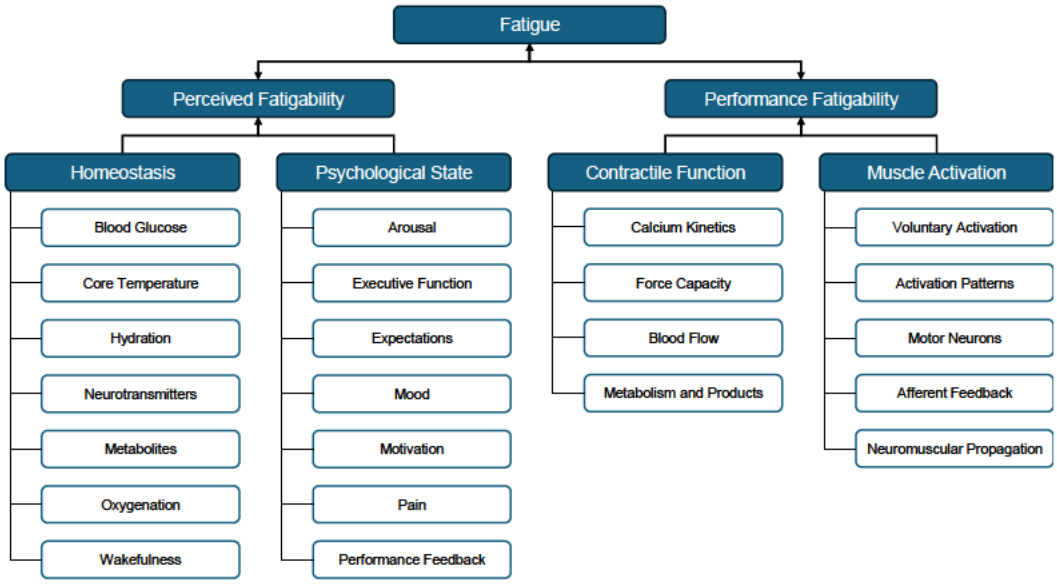


Fig. 2.1.: Taxonomy of fatigue factors by Enoka and Duchateau [110].

Based on this taxonomy first proposed by Enoka and Duchateau [109], Behrens et al. [34] presented an adapted motor and/or cognitive task-induced state fatigue framework with four interdependent dimensions and their respective determinants (see Figure 2.2).

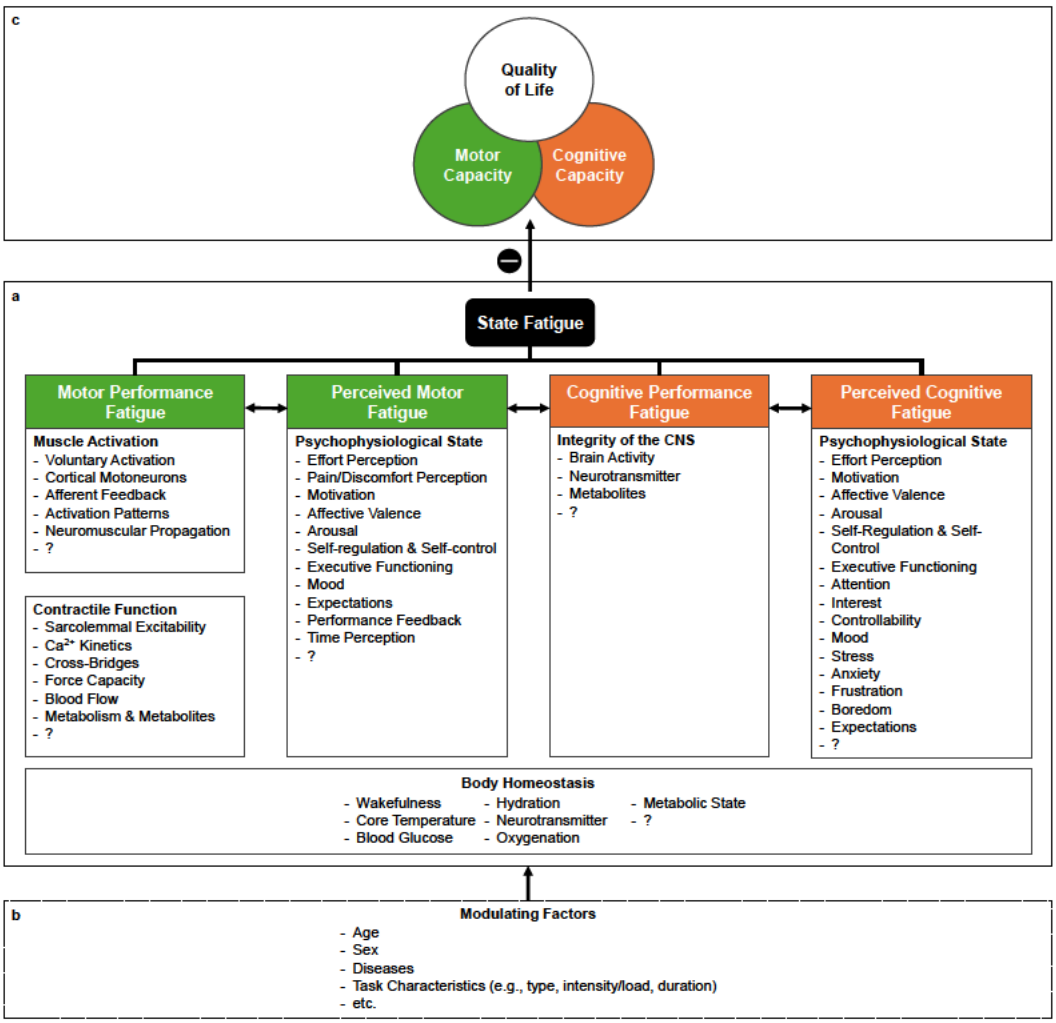


Fig. 2.2.: Taxonomy of fatigue factors by Behrens et al. [34].

As shown in Figure 2.1 and Figure 2.2, various possible factors contribute to fatigue, regardless of how they are categorised. Both taxonomies distinguish between the interdependent attributes of perceived fatigue and performance fatigue. Both fatigue attributes and their factors are interdependent and should not be considered in isolation. Consequently, there is no single factor that primarily determines performance fatigue and perceived fatigue in response to motor and cognitive tasks [34]. Performance fatigue depends on the contractile capabilities of the involved muscles and the capacity of the nervous system to provide an adequate activation signal for the task. Perceived fatigue is derived from the sensations that regulate the integrity of the performer based on the maintenance of homeostasis and the

psychological state of the individual [110]. Furthermore, Venhorst et al. [338] proposed a three-dimensional dynamic system framework to better understand the psychophysiological determinants of perceived motor fatigue. This system distinguishes three dimensions: sensory-discriminatory, affective-motivational, and cognitive-evaluative.

2.2.3 Fatigue Factors

Regardless of how fatigue factors are structured, fatigue should be understood as multi-factorial [110]. Evans et al. [114] hypothesised that fatigue can be compared to systems biology, where multiple components combine simultaneously in complex and dynamic interactions to produce emergent properties in organisms. This is in contrast to the reductionist perspective that a complex system can be fully understood by gaining knowledge of its isolated parts [114].

Fatigue is influenced by cardiovascular, respiratory, metabolic and neuromuscular factors and their interactions. In addition to the various physiological processes and their interactions, fatigue-related factors identified in fields other than exercise psychology, such as cognitive neuroscience and genetics, must not be ignored when attempting to understand exercise-induced fatigue [269].

A more detailed description of the various factors contributing to fatigue that are frequently mentioned in the literature can be found in the Appendix C. In summary, fatigue depends on several interdependent factors and there are several proposals on how to structure their interdependence. This is probably why there are many different approaches to measuring fatigue in the literature. Fatigue is quantified through its multiple effects, with different emphases in each discipline; fatigue is thus a multidimensional construct, studied through approaches that depend on the main interest of the research team and therefore with a limited focus [269].

2.2.4 Fatigue Measurement

Although fatigue is relevant in many research fields and domains, there is no standard for measuring fatigue. Whether measuring the cause, consequence, or subjective state, there is no clear signature of fatigue [269]. This is probably why there are such a wide variety of approaches to fatigue measurement. Regardless of the method, Behrens et al. [34] suggested that perceived motor fatigue and its contributing factors should be assessed before, during, and after fatiguing exercise. Motor fatigue is usually quantified as the decrease in peak force (torque) after an exercise intervention, although decreases in power, speed, or accuracy can also be measured [34, 199].

Halson [143] distinguished between external and internal load. External load quantifies task-related parameters (e.g., power output or speed) that are independent of individual characteristics. In contrast, internal load reflects the physiological and psychological load imposed by the task (e.g., heart rate variability or perceived exertion). Another approach, suggested by Goyal et al. [136], is to distinguish between subjective and objective measurement, which is adopted in this thesis.

Objective Fatigue Measurement

Objective measurement of fatigue is quantifiable, i.e., it is based on physiological, kinetic, or contextual data [269]. This measurement approach can reduce the possibility of self-deception, falsification, fabrication, attention, or recall bias that is usually present in subjective data collection [307]. Objective measurement can be further classified as either obtrusive or unobtrusive [23].

Obtrusive Fatigue Measurement Obtrusive measurement provides accurate quantitative data [23]. For example, in sports science, saliva or blood samples are often collected for cortisol, testosterone, creatine kinase, or lactate analysis [80, 108, 143]. Other potential markers include levels of blood glucose, muscle glycogen, blood acid, muscle acid, skeletal muscle ergoreceptors, blood oxygen levels, lipid stores, liver

glycogen stores, cardiovascular status, immune system activity, and body temperature [114]. During high-intensity exercise, large changes in metabolites and ions are observed within the working muscles; disturbances in the concentration of muscle lactate, hydrogen, potassium, and calcium ions are associated with fatigue and thus ionic regulation becomes critical for muscle membrane excitation, contraction, and energy metabolism [48].

There are also many non-invasive but obtrusive measurement methods for neural and muscular mechanisms, such as peripheral nerve stimulation, transcranial magnetic stimulation, structural magnetic resonance imaging, electromyography, positron emission tomography, electroencephalogram, magnetoencephalography, functional near-infrared spectroscopy, 31-phosphorus magnetic resonance spectroscopy, and electroencephalography [199, 34]. The cardio-respiratory approach is an obtrusive yet non-invasive method of detecting fatigue, commonly used in sports science. It requires the use of a face mask to measure the ability of the circulatory and respiratory systems to deliver oxygen. Other studies refer to this method as VO_2max, which stands for the maximum volume of oxygen consumption measured during incremental exercise [108].

The main drawback of most of these obtrusive measures is that they are usually not suitable for real time monitoring systems or frequent sampling due to bulky devices or the requirement for post-analysis in a laboratory [23]. Halson [143] also noted that the use of biochemical, hormonal, and immunological measures is not currently justified based on the limited research. These measures can be costly, time consuming, and impractical in an applied environment.

Unobtrusive Fatigue Measurement Unobtrusive measurement collects data either from wearable sensors attached to the body or from portable sensors placed in the immediate environment [23, 108]. In general, such unobtrusive measures can provide continuous data in real time [130].

Wearable sensors include, for example, *electrocardiography* (ECG) [4], blood flow [139], *photoplethysmogram* (PPG) [300], *inertial measurement unit* (IMU) [26], plantar pressure [20], *global positioning system* (GPS) [336], *surface electromyography* (sEMG) [69], *electrodermal activity* (EDA)[2] [252], *electroencephalography* (EEG), [357], *respiration* (RESP) [331], *skin temperature* (ST) [22], and *microphone* (MIC) [195]. Non-wearable but portable sensors include, for example, infrared image sensors (e.g., Kinect) [9], video cameras [341], eyelid activity sensors [364], thermal imaging [63], and force plates [67]. Many studies also use multiple wearable sensors, also referred to as a multi-sensor or multi-modal approach. In particular, Butkevičiūtė et al. [59] used multiple sensors (ECG, EMG, EEG, IMU) to detect different types of fatigue. A comprehensive survey of unobtrusive techniques for monitoring muscle fatigue can be found in Li et al. [229]. A review of sensors for detecting physical exertion with ML in the workplace can be found in Lambay et al. [217].

Wearable sensors in the form of smartphones, smartwatches or embedded systems tend to be low cost, widely available, potentially easy to use, and suitable for everyday use [130]. For example, smartwatches can monitor heart rate with clinically acceptable accuracy and could be considered safe for use in cardiac rehabilitation training programmes [108]. Smartphones typically consist of IMUs and video cameras, which are more suitable for automated data collection [143] than manual measurement of movement (e.g., by using a stopwatch or tape measure).

However, unobtrusive methods may not capture the perceived (motor and/or cognitive) fatigue (see Section 2.2.2). According to Ameli et al. [14], unobtrusive methods cannot provide comprehensive information on muscle fatigue due to their limited ability to record different aspects of movement. In addition, unobtrusively collected data may be susceptible to noise or artefacts due to poor sensor fixation and physical activity [130]. In addition, measures of training load such as power, work, energy, torque, or velocity are specific to the type of training, as the validity of

[2]Formerly known as *galvanic skin response* (GSR).

a measure depends on the context. For example, heart rate is a less valid measure of internal load for resistance training or short intermittent high intensity efforts than for endurance, long distance or interval training. Furthermore, a single measure may not have the same level of validity. For example, muscular fatigue increases both heart rate and perceived exertion, whereas mental fatigue only increases RPE [164].

Subjective Fatigue Measurement

Subjective fatigue measurement is traditionally used by psychologists in the form of questionnaires, interviews, or self-reports. Subjective measures are less suitable for frequent or real time monitoring of fatigue, but can still be used to determine fatigue before, during, and after a task [307]. A common approach is to define fatigue on the basis of exceeding a certain score on fatigue questionnaires [199].

Fatigue Scales Most clinical fatigue studies use self-report scales that can be broadly classified as measuring perceptions of fatigue [199]. Available scales vary widely in how they measure fatigue [217], including questions about momentary (state) perceptions, chronic characteristics (trait perceptions), the impact of fatigue on function, ratings of related constructs (e.g. tiredness), dimensions of fatigue (e.g. mental vs. physical), and severity [199]. Some scales have been developed for specific populations, but it is not clear whether such scales offer advantages over general scales [199]. Billones et al. [43] identified 23 different clinical measures used to assess fatigue in non-oncological conditions, with the Fatigue Severity Scale being the most commonly used clinical measure across different conditions.

The *ratings of perceived exertion* (RPE) scale[3] by Borg [50] is commonly used in sports science [34, 236, 113]. It is a subjective measure of an individual's perceived level of exertion during physical activity. The scale is a numerical rating system

[3]The similar Borg CR10 scale is derived from psychophysical scaling methods. It is more complex in its construction and is usually recommended for assessing pain rather than exertion [50].

ranging from 6 to 20, with most numbers corresponding to a verbal anchor, as shown in Table 2.2.

Tab. 2.2.: The RPE scale by Borg [50].

6	No exertion at all
7	Extremely light
8	
9	Very light
10	
11	Light
12	
13	Somewhat hard
14	
15	Hard (heavy)
16	
17	Very hard
18	
19	Extremely hard
20	Maximal exertion

As explained by Borg [50], studies using RPE should consider the five principles in its application (see a detailed description in Appendix F and G):

1. Briefly explain the importance of RPE inquiries during testing.
2. Provide comprehensive instructions on how to assess perceived effort.
3. Explain the RPE scale, including what to rate, how the scale works, and the meaning of verbal anchors.
4. Minimise distractions and external factors that may influence performance or RPE.
5. Establish a positive and collaborative relationship with the subject while maintaining standardised testing procedures. Adapt to individual personality factors and unexpected situations.

Correlation with Objective Measurement To date, there have been inconsistencies in finding significant correlations between objective measures with subjective fatigue questionnaire data [43, 215]. One reason is that subjective quantification is prone to recall errors [217]. Another reason is that sensory perception often does not grow

linearly with physical stimulation, but follows positively or negatively accelerating power functions [50].

Borg [50] argued that the RPE scale has high validity coefficients with heart rate and oxygen uptake. For example, an RPE of 13 would correspond approximately to a heart rate of 130 beats per minute, depending on factors such as age and physical condition. However, the validity may not be as high as previously thought [143]. On the other hand, the RPE scale has been shown to better represent a person's performance in practice than monitoring heart rate alone [108].

2.2.5 Summary

Fatigue is a multifaceted phenomenon that has been studied in various fields, including cognitive neuroscience, exercise physiology, psychology, and medicine. Despite its widespread relevance, there is still no universally accepted definition of fatigue, probably due to the different contexts in which it is used.

One of the key challenges in defining fatigue is its multidimensional nature. It encompasses both subjective perceptions, such as feeling tired, and objective changes in performance, such as a decline in muscle strength or mental acuity. Researchers have proposed various taxonomies to classify fatigue, distinguishing between factors such as perceived and performance fatigue, and further categorising it based on physiological, psychological, and cognitive determinants.

The methods used to measure fatigue also vary widely depending on the focus of the research. Subjective measures, such as self-reports and questionnaires like the Borg RPE scale, provide insight into how individuals perceive their fatigue, but can be influenced by individual bias. Objective measures can be either obtrusive (e.g., blood and saliva samples) or non-intrusive (e.g., wearable sensors) and can provide quantifiable data frequently through physiological, kinetic, or contextual indicators, but may only capture a limited aspect of fatigue.

2.3 Human Activity Recognition

This section provides a brief introduction to HAR[4], including its definition, main applications, existing taxonomies, and different sensor technologies. It is also shown why HAR is a key research area of this thesis, as HAR research covers not only human activities, but also the state of the body, such as fatigue.

2.3.1 HAR Definition

As with fatigue, there is no commonly agreed definition of HAR [57]. The following definition is adopted in this thesis: HAR is the ability to recognise or detect current human activity (or status) based on information received from various sensors. These sensors may include cameras, wearable sensors, sensors attached to objects of daily use, or deployed in the environment [162].

Human activity is a human behaviour in relation to the body or the environment. The detection of human activity aims to capture the action and/or status of an individual (agent) from a series of observations. A human activity can be either atomic or composed of many primitive actions performed in some sequential order [39]. Human activities can be categorised into a hierarchy of human activities according to their complexity, scaling from simple actions to more complex events [32, 91]: (1) elementary human actions such as bending an elbow, (2) gestures such as applause, (3) behaviour based on specific situations such as exercises, (4) interactions based on human to human such as shaking hands or human to object such as cooking, (5) group actions performed by a group of people such as cuddling, and (6) events that take place in specific environments such as weddings. More general taxonomies and categories of human activities can be found in [193, 267, 6].

[4] A related area of research is human action recognition, where human action is defined as an observable entity that can be decoded by another entity, including a computer, through various sensors [62].

2.3.2 HAR Applications

HAR is a multifaceted research field, covering almost all human activities. For this reason, HAR requires interdisciplinary knowledge to understand the behaviour and to provide proper assistance [39]. There are various application areas such as healthcare, smart environments, security and surveillance, human–computer interaction, indoor navigation, shopping experience, autonomous driving, human–robot interaction, smart home, and entertainment [162, 91].

2.3.3 HAR Taxonomies

A variety of taxonomies have been proposed for HAR [103, 339]. Hussain et al. [162] divided HAR into two main categories: vision-based and sensor-based, but the distinction is rather arbitrary as vision-based methods also incorporate sensors. On the other hand, Vrigkas et al. [339] classified HAR into unimodal and multimodal methods, but this classification does not clearly separate non-hybrid or physiological approaches. Bulling et al. [57] described HAR systems using five characteristic dimensions: execution (offline, online), generalisation (user-independent, user-specific, temporal), detection (continuous, isolated), activity types (periodic, sporadic, static), and system model (stateless, stateful).

Some researchers categorised HAR according to the object being tracked. For example, Hussain et al. [162] proposed three categories: action-based, motion-based, and interaction-based. Similarly, Vrigkas et al. [339] introduced six classes including gestures, atomic actions, human-to-object, or human-to-human interactions, group actions, behaviours, and events. However, such classifications do not cover body states. Another approach is to classify HAR according to the intended task [39].

Other taxonomies are based on sensor characteristics, such as active versus passive sensors, intrusive versus non-intrusive sensors, or by deployment method, including wearables, objects, or environmental sensors [39]. In addition, HAR is often categorised by the sensor technology like the sensor type, including inertial, pressure,

acoustic, vibration, ultrasonic, contact, electromagnetic, magnetic, visual, infrared, and radio frequency sensors. Another approach is based on the operational position of the sensor, such as environmental (e.g., pressure, temperature, humidity, or open/closed states), ambient (e.g., cameras, microphones, radio frequency, or motion detectors), and object attached sensors (e.g., inertial, heart rate, pulse, or electrical activity). A further classification refers to the underlying sensing principles, by organising sensors according to the types of waves they use (e.g., visible spectrum, infrared, radio frequency, mechanical waves, or vibrations) [103].

In this thesis, the following two HAR taxonomies are adopted because they focus not only on the detection of human activity, but also consider the perceived state of the individual (i.e., fatigue): one is based on the targeted task, the other on the physical measure of the sensor technology.

Targeted Task

Bian et al. [39] proposed a taxonomy organised into three classes according to the attributes of the targeted tasks: "Where" refers to body position-related services, including indoor positioning and tracking. "What" focuses on action-related recognition, including tasks such as fall detection and gait analysis. "How" focuses on the status of the body, including aspects such as emotion detection, stress sensing, and heart rate. Figure 2.3 illustrates this taxonomy and highlights the focus of this thesis, namely body status-related emotion and stress sensing (i.e., fatigue).

Physical Measure

The type of sensor used in HAR applications has a considerable influence on the performance and capabilities of the system. A key aspect is the need to balance the trade-offs between factors such as accuracy, computational resources, power consumption, and user acceptance. Each sensing technology has its own unique set

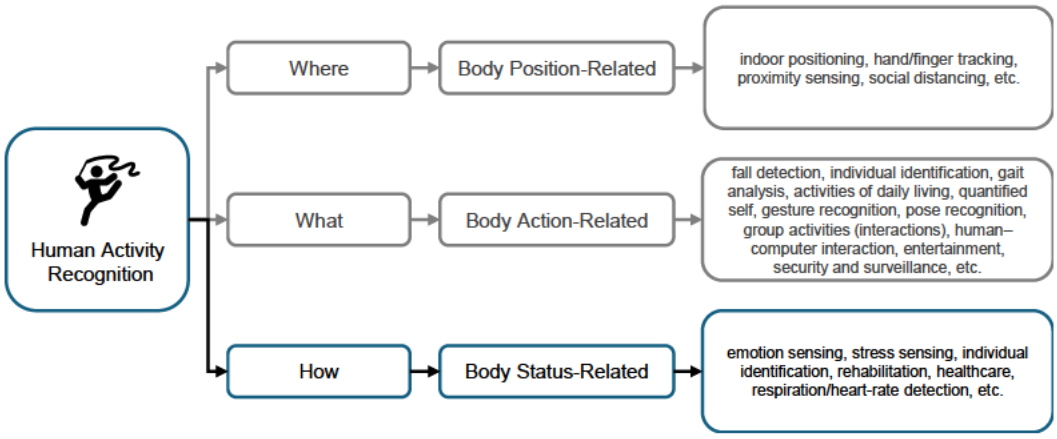


Fig. 2.3.: HAR taxonomy by Bian et al. [39]. The highlighted path represents the focus of this thesis.

of benefits and challenges, requiring careful evaluation of its suitability for specific applications and contexts [103].

According to Fu et al. [124], the same human action can be measured by different types of sensors, but the pool of actions is large, making action-based comparisons difficult. Therefore, they proposed a classification based on White [349], who used physical measures to facilitate comparison between sensors. Bian et al. [39] refined these works and proposed the following five classes based on sensing techniques: field, mechanical kinematic, physiological, wave, and hybrid/other. Figure 2.4 illustrates this taxonomy and also highlights the focus of the case study in this thesis described in Chapter 4. A description of each sensing technique can be found in Appendix J.

2.3.4 Summary

HAR is defined as the ability to recognise or detect current human activity (or status) based on information received from various sensors. Human activity is a human behaviour in relation to the body or the environment. The detection of human activity aims to capture the action and/or status of an individual (agent) from a series of observations. The applications for HAR are diverse, spanning areas

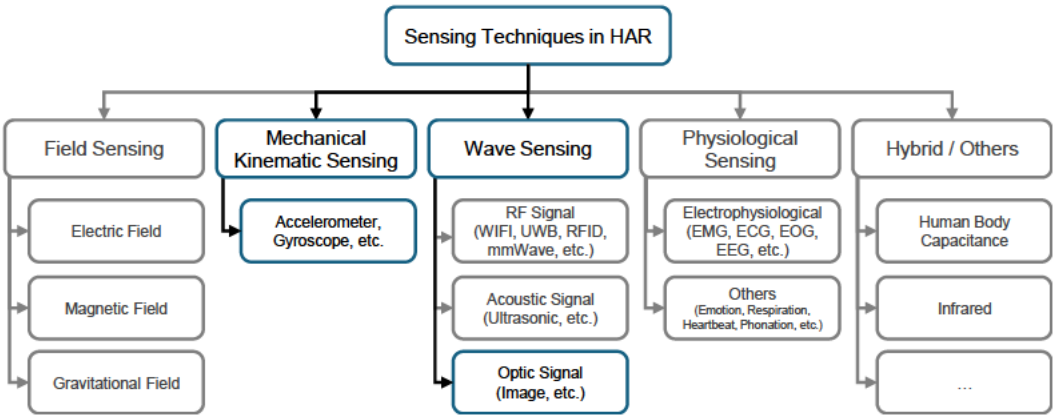


Fig. 2.4.: HAR sensing techniques by Bian et al. [39] based on [349, 124]. The highlighted paths represents the focus of the case study in this thesis.

such as healthcare, shopping experience, and smart environments. There are many taxonomies of HAR based on different criteria, such as sensor types, tracked objects, or deployment methods. This thesis adopts two specific taxonomies for HAR, both incorporating the state (i.e., fatigue) of the body: one is focused on the targeted task (where, what, why) and the other on the sensing technology (field, mechanical kinematic, physiological, wave, and hybrid/other).

2.4 Survey of Related Works

This section reviews the primary studies on HAR and exercise fatigue detection using a systematic approach as described by Lame [218] to create a literature survey:

1. Aim: To find common methods and techniques used by related works on sensor-based fatigue detection using ML to handle small data and improve generalisability.
2. Inclusion and Exclusion Criteria: Publications on sports and physical work activities, such as material handling, are included because the fatigue detection methods are the same in both domains. However, publications on the detection of fatigue in animals are excluded, although the methods are basically the same as in humans, see for example Darbandi et al. [94], however, subjective

questionnaires are not possible in such studies. Studies that do not use any type of ML are also excluded, as are publications that only describe an intention for future research.

3. Quality Assessment: Publications must be peer-reviewed.
4. Locate Publications: Find related works in human-centred computing, sports science, and sports medicine databases using predefined search terms.
5. Select Publications: First, titles and abstracts are screened. Second, the full texts that were not excluded in the first step. In addition, a backward and forward search is performed on each publication that meets the inclusion criteria (see Locate Publications). This step is detailed in the following Section 2.4.1.
6. Data Extraction: Relevant information from the selected publications is extracted, such as number of samples, evaluation and augmentation methods, and statements about generalisability (see Appendix B).
7. Analysis: A statistical summary of the extracted information is created resulting in a literature survey (see Section 2.4.2).
8. Interpretation: The results of the survey, including the findings of the case study, are discussed in Chapter 6.

The findings extracted from the literature serve as the foundation for the Fatigue Recognition Chain framework in Chapter 3 and the case study in Chapter 4. As the case study of this thesis focuses on squats, studies related to squat-based fatigue detection are examined in a separate section (see Section 2.4.3).

2.4.1 Selection Procedure

A literature review was conducted to provide an overview of common techniques applied to fatigue detection during physical activity. The following search term was used in the academic search engines IEEE Xplore, Scopus, and PubMed NIH for publications in the last 15 years:

```
(fatigue OR exertion) AND
(sport OR sports OR training OR exercise) AND
(sensor OR sensing OR signal OR signals OR wearable OR wearables) AND
("machine learning" OR "ML")
NOT virtual NOT creativity NOT security NOT driver NOT drowsiness NOT "mental fatigue"
```

As the ACM Digital Library returned a large number of results, the first line of the search term was adjusted and only publications in the last 5 years were included:

```
("fatigue recognition" OR "fatigue detection" OR "fatigue prediction" OR exertion)
AND ...
```

ACM provided 4042, IEEE 175, Scopus 295, and PubMed 115 results based on a search on 02 September 2024. Figure 2.5 shows the flow diagram of the literature review. For each search engine, the titles and abstracts of the search results were

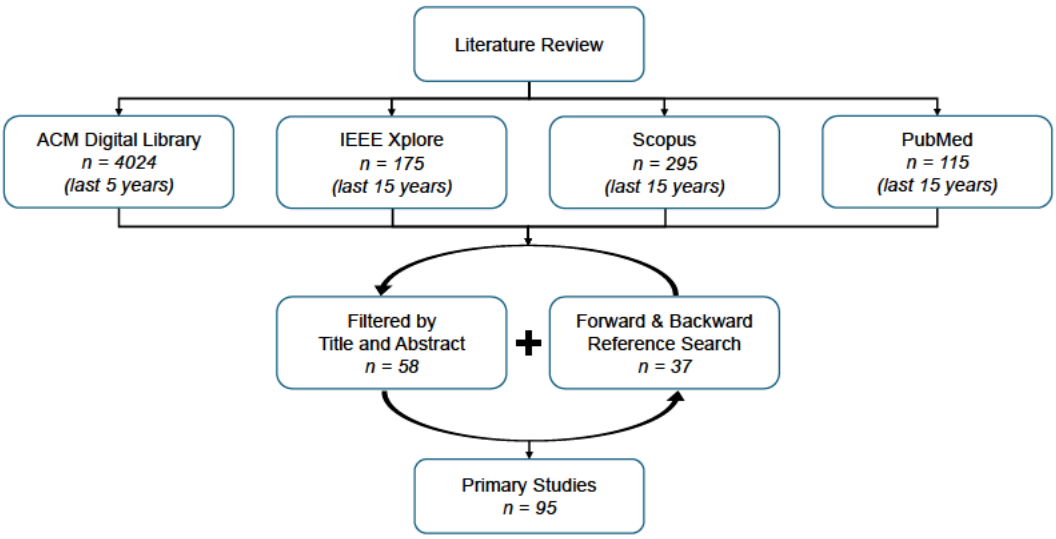


Fig. 2.5.: The flow diagram of the literature review.

reviewed to determine whether any type of sensor was used to detect fatigue and whether fatigue was caused by a specific physical activity.

2.4.2 Related Works

A total of 95 related works were identified, as shown in Table 2.3 (more detailed information can be found in Appendix B). These related works recruited different

numbers of subjects, ranging from 1 to 80, with an average of 21.1 (median: 17) subjects. Other reviews on fatigue detection during physical activity found similar numbers. Martins et al. [248] conducted a broad review of fatigue not limited to ML and exercise, with varying numbers of subjects from 3 to 50 and an average of 14 subjects. Marotta et al. [247] reviewed studies on fatigue detection with accelerometers located at lower limbs during cycling exercise, ranging from 3 to 222 subjects per study with an average of 23.1 subjects. Sun et al. [320] analysed studies on exercise fatigue detection with sEMG sensors, ranging from 6 to 58 subjects and an average of 28.1 subjects.

Tab. 2.3.: Overview of the related works.

Authors	Year	n	Exercise	Ground Truth	Classes	Samples	Best Acc. (%)
Albert and Arnrich [9]	2024	16	Squats	RPE20	14	2304	-
Gan et al. [126]	2024	16	Squats	RPE10	3	31	80.7
Huang et al. [161]	2024	7	Static	RPE10	2	N/A	86.13
Ma and Guo [240]	2024	30	Yoga	Blood samples	N/A	N/A	N/A
Mu et al. [256]	2024	32	Running	Visual Analog Scale	2	2560	94
Wang et al. [342]	2024	32	Bicep Curls	K-means clustering	3	580	83.3
Wang et al. [346]	2024	5	Material Handling	Questionnaire	4 (5)	483 (12558)	94.7
Yao [358]	2024	20	N/A	K-means clustering	2 (3)	N/A	83.11
Zhang et al. [365]	2024	18	Running	First vs last 5 min	2	5400	99
Adapa et al. [1]	2023	11	Bicep Curls	Activity Intensity	2	N/A	86.51
Antwi-Afari et al. [20]	2023	10	Material Handling	RPE20	2, 3, 4	1289	96.9
Anwer et al. [21]	2023	15	Material Handling	RPE20	4	1425	93.5
Biró et al. [45]	2023	9	Cycling, Running, Football	RPE20, Heart Rate	N/A	N/A	90
Biró et al. [44]	2023	19	Running	Activity Intensity (Beep test)	2	1201	59
Bouteraa et al. [52]	2023	57	Wrist Torque	Uncertainty algorithm	2	N/A	92.62
Cañellas et al. [63]	2023	80	N/A	Linearly annotated	101	418813	-
Concha-Pérez et al. [83]	2023	30	Squeeze/Release (Arm)	Activity Intensity	2	N/A	95.7
Dang et al. [92]	2023	10	Dynamometer	Activity Intensity	3	N/A	93.5
De Vito et al. [97]	2023	1	Material Handling	N/A	2	5634	83.9
Dimmick et al. [102]	2023	16, 9	Running	RPE, MLSS, first and last km	2	N/A	68.9
Feng et al. [116]	2023	25	Rope-Skipping	Activity Intensity	2	N/A	-
Kathirgamanathan et al. [191]	2023	19	Running	Activity Intensity (Beep test)	2	5510	97, 59
Liu et al. [235]	2023	20	Elbow	RPE20	4	7560	96.67
Marena et al. [246]	2023	5	Material Handling	Metabolic rate	N/A	N/A	-
Perpetuini et al. [275]	2023	10	Squats	Activity Intensity	2	N/A	-
Pirscoveanu and Oliveira [278]	2023	43	Running	RPE20	14	N/A	-
Pravin et al. [280]	2023	N/A	Bicep Curls	Activity Intensity	2	24	87.5
Smiley et al. [313]	2023	10	Cycling	RPE10	2	150	80
Valla et al. [335]	2023	41	Archimedean Spiral Test	Questionnaire	2	33	78.8
Albert et al. [10]	2022	12	Squats	RPE20, lactate	14	N/A	-
Bustos et al. [58]	2022	24	Running	RPE20	4	750	88
Jaiswal et al. [168]	2022	32	Walking	First sets vs last two sets	2	N/A	80.5
Umer et al. [332]	2022	10	Material Handling	RPE20	14, 4	1286	64.2, 75.73
Cheah et al. [69]	2022	4	Sit-Ups	First vs last 20% reps	2	1092	65.3
Escobar-Linero et al. [112]	2022	7	Material Handling	RPE20	4	360	91
Guo et al. [139]	2022	10	Bicep Curls	RPE	3	800	92
Jiang et al. [176]	2022	12	Squats	RPE10	10	N/A	83.7
Li and Chen [225]	2022	20	Pilates	RPE	3	1200	94.25
Shi et al. [306]	2022	10	Walking	Activity Intensity	5	N/A	88.9
Triantafyllopoulos et al. [327]	2022	48	Running	RPE20	14	N/A	-
Wang et al. [343]	2022	19	Running	RPE20	3	N/A	91.1
Zhu et al. [374]	2022	24	Walking, Cycling, Running	Activity Intensity	6	14400	97.7
Chen et al. [71]	2021	40	Material Handling	Control group	2	80	72

Table 2.3 continued from the previous page

Authors	Year	n	Exercise	Ground Truth	Classes	Samples	Best Acc. (%)
Chen et al. [70]	2021	47	Material Handling	Control group	2	94	89.47
K et al. [181]	2021	58	Bicep Curls	First vs last rep	2	116	94.04
Sadat-Mohammadi et al. [288]	2021	15	Material Handling	Activity Intensity, NASA-TLX	3	N/A	93.4
Wang et al. [344]	2021	20	Cycling	Ventilation Threshold	2	8872	95.15
Zhang et al. [370]	2021	2, 10	Shoulder	10 s after exhaustion	2	18740, 93998	96.45; 78.25
Aguirre et al. [4]	2021	60	Sit-to-Stand	RPE10	3	660	83.2
Balaskas and Siozios [28]	2021	14	Running	Clustering	2	N/A	43
Chalitsios et al. [67]	2021	13	Running	Ventilatory Threshold	2	29650	91.4
Chen et al. [72]	2021	10	Dumbbell (pick-up)	Manually labelled	2	5000	90.4
Elshafei et al. [108]	2021	20	Bicep Curls	RPE20	2	3000	18–95
Guan et al. [138]	2021	14	Running	RPE20	3	N/A	80.6
Jiang et al. [175]	2021	14	Squats, Jacks, Touch	RPE10	10	1790, 1240, 1140	-
Karvekar et al. [189]	2021	24	Squats, Walking	RPE20	2, 3, 4	1240, 1800, 2400	91, 78, 64
Kuschan and Krüger [212]	2021	9	Material Handling	RPE10	3, 5	282	83.8, 80.9
Lambay et al. [216]	2021	24	Material Handling	RPE	2	N/A	65
Wang and He [341]	2021	12	Running	RPE20	4	6624	87.7
Davidson et al. [95]	2020	12	Running	RPE20	2	112	84.8
Luo et al. [238]	2020	27	Daily Activities	Fatigue Assessment Scale	2	N/A	71.4
Umer et al. [331]	2020	10	Material Handling	RPE20, SWAT	14	1286	98.5, 95.3
Wang et al. [345]	2020	20	Cycling	Ventilation Threshold	2	100	83.51
Guaitolini et al. [137]	2020	13	Walking, Running	First vs other reps	2	26	84.6
Maman et al. [242]	2020	15	Material Handling	RPE	2	234 (46800)	85
Nasirzadeh et al. [261]	2020	8	Material Handling	RPE20	2	3456, 1728, 691	90.36
Sani et al. [293]	2020	8	Material Handling	RPE	2	N/A	78.2
Zhang and Wang [364]	2020	20	Ball Sports	PERCLOS P80	2	8000	90
Chowdhury et al. [77]	2019	22	Walking, Running	RPE20	3	615	-
Geurkink et al. [129]	2019	46	Football	RPE10	10	913	91.7
Jebelli et al. [171]	2019	10	Material Handling	Activity Intensity	2, 3	N/A	90, 87
Karvekar et al. [188]	2019	24	Squats, Walking	RPE20	2, 4	N/A	91, 61
Papakostas et al. [265]	2019	10	Shoulder	Exhaustion plus 10 s	2	90	-
Yang and Ren [357]	2019	20	Muscle Chair	RPE10	2	220*	90
Wu et al. [353]	2018	N/A	Running, Walking, Pedalling	Activity Intensity	2	148	98.65
Baghdadi et al. [26]	2018	20	Material Handling	RPE20	2	1000	90
Beéck et al. [33]	2018	29	Running	RPE20	14	7607	-
Gordienko et al. [135]	2018	N/A	Walking, Running, Skiing	Clustering	N/A	N/A	-
Jamaluddin et al. [169]	2018	20	Running	Questionnaire	2	N/A	98
Karthick et al. [187]	2018	52	Bicep Curls	First segments vs last segment	2	N/A	91.5
Aryal et al. [22]	2017	12	Material Handling	RPE20	4	253	82.6
Lopez et al. [237]	2017	19	Running (stairs)	Activity Intensity	2	5700	81.51
Shahmoradi et al. [302]	2017	6	Reaching (arm)	Max. Voluntary Contract.	3	N/A	95.3
Vandewiele et al. [336]	2017	45	Football	RPE10	10	913	-
Buckley et al. [55]	2017	21	Running	Last 400 m	2	584	75
Maman et al. [243]	2017	8	Material Handling	RPE20	2	144	-
Carey et al. [65]	2016	45	Football	RPE10	10, 15	3398	-
Kupschick et al. [211]	2016	22	Material Handling	RPE20	2	533	85.8
Pernek et al. [274]	2015	11	Dumbbell (upper body)	RPE20	14	264	-
Bilgin et al. [41]	2015	31	Running	Bruce protocol	2	N/A	92
Karg et al. [186]	2014	7	Squats	Questionnaire	5	445	81
Zhang et al. [367]	2013	17	Squats, Walking	Until 60% maximal exertion	2	340	90
Janssen et al. [170]	2011	9	Leg, Walking	Activity Intensity	2, 3	162	98.1
Subasi and Kiymik [319]	2009	14	Dumbbell	N/A	2	1100	91
Karg et al. [185]	2008	14	Rowing, Walking	Activity Intensity	2	N/A	100
This thesis	2024	48	Squats	RPE20	2, 3, 4	3595	78.4

The related works were published between 2008 and 2024 (median: 2021), as shown in Table 2.4.

Tab. 2.4.: Number of related works per year.

2024	2023	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008
7	20	13	17	9	6	6	6	2	2	1	1	0	1	0	1	1

Physical Activities

Table 2.5 lists the physical activities used in related works. 72 related works used sports training as an activity, while 21 used work-related material handling. Running was the most frequently used exercise, followed by material handling, walking, and squats. At least two different activities were used in 10 related works.

Tab. 2.5.: Physical activities and the number of related works.

Activity	Running	Material Handling	Walking/Gait	Squats	Bicep Curls	Cycling	Football	Dumbbell	Other	N/A
Count	25	21	12	10	7	5	4	3	8	2

Note: 24 related works utilised multiple different activities.

Sensors

Table 2.6 lists the most frequently used sensors in the related works. ECG, IMU, and sEMG were the most utilised sensors for exercise fatigue detection. 45 related works utilised multiple different sensors, and 44 only a single sensor, with IMU counted as a single sensor.

Tab. 2.6.: Sensors and the number of related works.

Sensor	ECG	IMU	sEMG	RESP	ST	GPS	Kinect	PPG	MoCap	EDA	FP	Thermal	Cam	Other	N/A
Count	38	37	26	11	8	5	5	5	5	4	4	3	1	8	1

Note: 45 related works utilised multiple different sensors.

Ground Truth

44 related works (46.32%) applied any form of the RPE scale (RPE20 or CR10) as ground truth.

A time-based approach was most commonly used by 26 of the 95 related works, where the ground truth was collected in time intervals (see also Table B.2 in the Appendix). The intervals range from 30 seconds to 24 h, with a median of 5 minutes. Table 2.7 lists these intervals. 8 related works collected the ground truth after a

Tab. 2.7.: Ground truth time intervals and the number of related works.

Interval	0.5 s	1 min	2 min	3 min	4 min	5 min	10 min	15 min	60 min	24 h
Count	2	2	4	1	1	5	7	1	1	2

certain number of exercise repetitions. The number of repetitions ranges from 5 to 15, with a median of 11 repetitions. Table 2.8 lists the number of repetitions used in the related works. In 6 related works, ground truth was assessed after a specific distance, ranging from 100 to 1000 m with a median of 400 m (typically 1 lap outdoors).

Tab. 2.8.: Ground truth repetition count and the number of related works.

Repetitions	**5**	**10**	**12**	**15**
Count	3	1	2	2

Table 2.9 shows the different ground truth approaches and how often they have been used in the related works. "Activity Intensity" means that multiple sessions of an exercise at different intensities are conducted to label the collected data. "First vs Last" means that the collected data was divided into certain time periods and the later periods were labelled as fatigue. Some studies utilised other subjective questionnaires than RPE. Clustering was applied in studies that applied deep learning methods. "Ventilation Threshold" was mainly used for cycling exercises. Other approaches were used only once or twice, including automatic linear annotation, Bruce protocol, control group fatigue assessment scale, lactate/blood samples, maximum voluntary contractions, metabolic rate, MLSS, PERCLOS-P80, SWAT, uncertainty algorithm, visual analog scale, and x seconds after exhaustion.

Tab. 2.9.: Ground truth approaches and the number of related works.

Approach	**RPE**	**Activity Intensity**	**First vs Last**	**Questionnaire**	**Clustering**	**Ventilation Thres.**	**Other**	**N/A**
Count	44	16	7	6	4	2	14	2

Note: 5 related works utilised multiple different ground truth approaches.

Samples

The number of samples collected by the related works ranges from 24 to 418813 with an average of 8759.9 and a median of 1046 samples per study (see Table B.4 in the Appendix). By far the most samples were collected by Cañellas et al. [63], who used an algorithm to automatically label video frames. There is no information on how many samples were collected in 33 studies.

Classes

Table 2.10 shows the number of classes (or labels) used in the related works. Each class requires a certain number of data points to train an ML model – the more classes, the more data is required. Binary classification is the most common, used by more than half of the related works. As RPE consists of 14 classes, different strategies are applied to reduce them. Of the 44 studies that used RPE, 31 studies reduced the number of classes for ML, usually by combining several classes and their samples into one class (group).

Tab. 2.10.: Classes (labels) and the number of related works.

Classes	**2**	**3**	**4**	**5**	**6**	**10**	**14**	**15**	**101**	**N/A**
Count	55	16	11	3	1	5	7	1	1	4

Note: 7 related works utilised multiple class counts.

Some related works reduced the number of classes by merging them. Table 2.11 shows different thresholds for merging classes into smaller numbers of classes. In general, these thresholds are less applicable to other research projects as they depend on the particular research design, sample selection, and class distribution.

Tab. 2.11.: Label reduction in the related works.

Authors	**Exercise**	**2 classes**	**3 classes**	**4 classes**
Elshafei et al. [108]	Bicep Curls	6–16 / 17–20	-	-
Smiley et al. [313]	Cycling	1–3 / 4–10	-	-
Liu et al. [235]	Elbow	-	-	6–12 / 13–16 / 17–18 / 19–20
Umer et al. [332]	Material Handling	-	-	6–10 / 11–13 / 14–16 / 17–20
Anwer et al. [21]	Material Handling	-	-	6 / 7–11 / 12–16 / 17–20
Aryal et al. [22]	Material Handling	-	-	6–11 / 12–14 / 15–16 / 17–20
Chen et al. [72]	Material Handling	6–15 / 16–20	-	-
Kupschick et al. [211]	Material Handling	6–10|12 / 15|13–20	-	-
Maman et al. [243]	Material Handling	6–12|14 / 13|15–20	-	-
Maman et al. [242]	Material Handling	6–12 / 13–20	-	-
Nasirzadeh et al. [261]	Material Handling	6–14 / 15–20	-	-
Antwi-Afari et al. [20]	Material Handling	-	-	6 / 7–11 / 12–16 / 17–20
Bustos et al. [58]	Running	-	-	6–11 / 12–14 / 15–16 / 17–20
Guan et al. [138]	Running	-	6–11 / 12–16 / 17–20	-
Mu et al. [256]	Running	-	1–4 / 4–7 / 8–10	-
Chowdhury et al. [77]	Running, Walking	-	6–11 / 12–14 / 15–20	-
Gan et al. [126]	Squats	-	0–1 / 2–4 / 5–10	-
Karvekar et al. [188]	Squats, Walking	6–16 / 17–20	-	-
Karvekar et al. [188]	Squats, Walking	6–14 / 15–20	-	6 / 7–10 / 11–15 / 15–20 [sic!]
Karvekar et al. [189]	Squats, Walking	6–14 / 15–20	6 / 7–13 / 13–20 [sic!]	6–11 / 11–13 / 13–15 / 15–20 [sic!]
This thesis	Squats	6–14 / 15–20	6–9 / 10–15 / 16–20	6–9 / 10–11 / 12–15 / 16–20

Imbalanced Classes

45 of the 95 related works reported if classes were balanced or imbalanced, as shown in see Table 2.12 (see more details in Table B.6 in the Appendix). Furthermore, 70 related works made one or two direct or indirect remarks on class imbalances. To handle imbalanced classes, 10 related works applied oversampling and 10 applied undersampling, with 2 related works using both over- and undersampling in the same study. However, the number or ratio of augmented samples was only reported by Wang et al. [346]. Some related works reported, how imbalanced data was handled. For example, Jiang et al. [176] and Maman et al. [242] duplicated samples (bootstrap). Guan et al. [138] applied SMOTE to increase the data of the minority classes. Aguirre et al. [4] mapped each class to the five closest repetitions for each subject. Jiang et al. [176] divided the subjects into fast and slow fatiguing subgroups according to the number of repetitions they have conducted prior to exhaustion. Baghdadi et al. [26] extracted data from the first and last 10 minutes to obtain an equal number of strides in both fatigued and non-fatigued states.

Tab. 2.12.: Imbalanced classes and data augmentation reported by the related works.

	Imbalanced Classes	Oversampled	Undersampled
Yes	23	10	10
No	22	85	85
N/A	46	2	2

Evaluation Types

Wang and He [341] identified the following three types to evaluate the performance of ML models:

- *Type 1 - Single Subject* (*T1-SOLO*): data from one individual is used for ML (e.g., if multiple data sets of one individual exist).
- *Type 2 - Leave No Subject Out* (*T2-LNSO*): Some data from all subjects is used as test set and excluded from the training set.
- *Type 3 - Leave One Subject Out* (*T3-LOSO*): Data from one subject is used as test set and excluded from the training set [37].

Depending on the total number of subjects, a single subject for evaluation usually results in a low testing rate. To mitigate this problem, a fourth type is proposed in this thesis to reduce the dependence of the evaluation on a single subject and to increase the size of the test set relative to the training set (see [41, 242]):

- *Type 4 - Leave Multiple Subjects Out* (*T4-LMSO*): Data from multiple subjects is used as the test set and excluded from the training set.

In addition, an ensemble approach can be implemented to train multiple ML models, each with a different test set: the final decision is then made, for example, by majority voting among all the models, which can reduce the dependence on a specific test set or subject(s) [26]. Furthermore, Adapa et al. [1] introduced leave-one-activity-out to examine the influence of different training loads and postures for five different exercises.

Table 2.13 shows by how many related works each evaluation type was used, with 13 related works that applied multiple evaluation types (more details can be found in Table B.4 in the Appendix).

Tab. 2.13.: Evaluation types and cross-validation (CV) and the number of related works.

Type	***T1-SOLO***	***T2-LNSO***	***T3-LOSO***	***T4-LMSO***	**CV**
Count	9	47	34	6	63

Note: 13 related works applied multiple evaluation types.

Cross-Validation

The evaluation is usually combined with cross-validation where the training set is divided into k folds. In each iteration, the ML model is then trained on k-1 folds and validated on the unknown remaining validation fold. The partitioning can be such that either all possible permutations are tested, or only a limited number of folds. Cross-validation can also be performed on the test set. For example in *T3-LOSO*, the test subject is swapped with another subject and the trained ML model is tested again until all subjects (or a limited, random number) have been the test subject once. Then, either the result of the ML model with the best score is taken [138] or an

average of all the tests subjects is calculated [265, 357]. When *T4-LMSO* is combined with cross-validation, multiple subjects are selected as the test set and multiple test sets are created for cross-validation. To reduce the number of test sets, only a limited number of (random) test sets may be used (Monte Carlo cross-validation [225]).

Cross-validation was applied in 63 related; and for 30 works, it is unknown whether cross-validation was used (see Table B.4 in the Appendix). Table 2.14 shows the number of folds used for cross-validation in the related works, with 5 folds being the most common, followed by 10 folds.

Tab. 2.14.: Folds for cross-validation and the number of related works.

Folds	**5**	**10**	**6**	**3**
Count	22	19	3	1

ML Models

Table 2.15 lists the ML methods commonly used in the related works, which is consistent (except for GANs) with the findings of Hooda et al. [157], who reviewed fatigue detection approaches using ML. On average, 3.4 models were trained by the related works, ranging from 1 to 14 models. 14 studies trained one ML model, while

Tab. 2.15.: ML methods and the number of related works.

Abbr.	**Full name**	**Count**
SVM	Support Vector Machine	57
RF	Random Forest	37
ANN	Artificial Neural Network	26
k-NN	*k*-Nearest Neighbours	25
LR	Logistic Regression	24
DT	Decision Trees	20
CNN	Convolutional Neural Network	17
LSTM	Long short-term Memory	13
NB	Naive Bayes	12
SVR	Support Vector Regression	5
LDA	Linear Discriminant Analysis	4
RNN	Recurrent Neural Network	4

Note: Other methods with less than four occurrences are not included.

26 studies trained at least two or more different models for comparison. The most commonly applied ML method was SVM, followed by RF, ANN, *k*-NN, LR, DT, CNN,

and LSTM. Most methods were used for supervised ML. Some works also adopted less common ML methods, for example Gradient Boosting Regressor, Hidden Markov Model, Gradient Descent, Gradient Boosting, Bagged Trees, Ensembles, and more.

The number of features used to train the ML models ranges from 1 to 531 features, with an average of 34.8 and a median of 13.5 features per work. For 7 related works the number of features was not available (see Table B.4 in the Appendix).

Reported Results

79 related works performed classification and 15 regression to evaluate the ML models, with 2 doing both. The best accuracy results that were reported in the related works average 85.7% with a median of 89.2%, ranging from 43% to 100%. Table 2.16 lists the best results and measures reported by the related works.

Tab. 2.16.: Results and measures reported by the related works.

Measure	Acc	Spec	Recall	Prec	F_1	R^2	RMSE	MAE	MAPE	CM
Mean	85.7	87.0	86.7	87.2	84.1	74.4	1.9	3.2	9.6	-
Median	89.2	92.0	89.8	90.1	85.2	86.0	0.9	1.8	9.6	-
Min	43.0	25.0	44.3	59.0	59.0	48.0	0.3	0.2	7.7	-
Max	100	100	100	99.0	97.3	89.3	13.6	16.5	11.5	-
Count	77	14	28	17	29	4	12	7	3	34

Accuracy in % (Acc), *Specificity in %* (Spec), *Recall in %* (Recall), *Precision in %* (Prec), *F_1 score in %* (F_1), *coefficient of determination in %* (R^2), *root mean square error* (RMSE), *mean absolute error* (MAE), *mean absolute percentage error* (MAPE), *confusion matrix* (CM)
Note: 65 related works applied multiple measures. Moreover, the measures Area Under the Curve and Receiver Operating Curve were used by some related works but not included in this survey.

21 of the 95 related works utilised accuracy alone to evaluate the performance of the ML models. Furthermore, 9 of the 79 related works, that used accuracy and other measures, reported less than 75% accuracy as the best result [44, 102, 55, 238, 216, 28, 71, 69, 332]. The lowest results were reported by Balaskas and Siozios [28] with 43%, while Umer et al. [332] aimed for a minimalist ML approach with as few input features as possible and reported 64.2% accuracy for the group model and 75.73% for the personalised model. 29 related works used F_1 scores [294], but these rarely specified the type (micro, macro, weighted).

Table 2.17 shows the sizes of the training and test sets typically chosen by the related works to train the ML models. The test ratio ranges from 50% to 95%, with an average of 17.6% and a median of 15%.

Tab. 2.17.: Size of the training set, test set, and samples in the related works.

Size	Training Set	Test Set	Samples
Mean	78.9%	17.6%	8759.9
Median	80%	15%	1046
Min	50%	2%	24
Max	95%	45%	418813

Generalisation

70 related works (77.77%) addressed in one or two sentences generalisation, overfitting, small data, and variability. 25 related works did not mention any of theses topics (see Table 2.18).

Tab. 2.18.: Number of related works that addressed certain topics.

Topic	Generalise	Small Data	Overfitting	Variability	None
Count	24	19	16	16	25

Some related works described specific findings or measures in regard to the four topics, they are highlighted in the following (see Table B.7 in the Appendix for details). Pernek et al. [274] was the only study among the related works that investigated how the ML models performed with different numbers of subjects. Baghdadi et al. [26] evaluated the ML performance by adjusting the size of the training and test sets. Some related works found significant intra- and inter-individual changes in the parameters between normal and fatigued states [185, 170, 265, 364, 331, 102, 44, 336]. Karg et al. [186] found no correlation between variance and fatigue; they suggested regression for small data. Lopez et al. [237] suggested transfer learning to deal with small data. Chowdhury et al. [77] specifically mentioned merging classes to improve generalisability. Some related works observed that adjacent data samples of fatigued states were prone to confusion during classification [235, 302, 211]. Guaitolini et al. [137] suggested that part of the misclassification error is related to population variability. Janssen et al. [170] selected subjects to create a

relatively homogeneous group in which subjects fatigued with comparable levels of exercise. Kathirgamanathan et al. [191] grouped similar subjects into separate data sets. Wang et al. [346] used data augmentation to improve generalisation, increasing the number of samples from 483 to 12558 by adding noise to the original data. Mu et al. [256] developed a custom transformer framework for learning small time series.

2.4.3 Related Works with Squats

This section takes a closer look at related works that have used squats as a physical exercise, since squats were also chosen for the case study of this thesis. 10 of the 95 related works used squats. As shown in Table 2.19 , these 10 works enrolled an average of 15.2 subjects, ranging from 7 to 24 subjects. The most commonly used

Tab. 2.19.: Overview of the related works that use squats as a physical exercise.

Authors	Year	n	Sensor	Ground Truth	Frequency	Classes	Samples
Albert and Arnrich [9]	2024	14	Kinect, ECG, IMU, kMeter	RPE20	12 reps	16	2304
Gan et al. [126]	2024	16	ECG	RPE10	30/15 reps	3	31
Perpetuini et al. [275]	2023	10	sEMG, Thermal	Activity Intensity	-	2	N/A
Albert et al. [10]	2022	12	IMU, ECG, Kinect, kMeter	RPE20, lactate	12 reps	14	N/A
Jiang et al. [176]	2022	12	IMUs	RPE10	5 reps	10	N/A
Jiang et al. [175]	2021	14	IMUs, FP, MoCap	RPE10	5 reps	10	1790, 1240, 1140
Karvekar et al. [189]	2021	24	IMU	RPE20	2 min	2, 3, 4	1240, 1800, 2400
Karvekar et al. [188]	2019	24	IMU	RPE20	2 min	2, 4	N/A
Karg et al. [186]	2014	7	MoCap	Questionnaire	5 reps	5	445
Zhang et al. [367]	2013	17	IMUs, FP, MoCap	60% max. exertion	Per set	2	340
This thesis	2024	48	IMU, PE	RPE20	10 s	2, 3, 4	3595

Note: Some related works [175, 189] evaluated the ML models with a different number of classes (2, 3 or 4), hence the different number of samples.

sensors were IMU (6x), ECG (3x), *Motion Capturing* (MoCap) (3x), Kinect (2x), *force plate* (FP) (2x), and EMG (1x). The most commonly used ground truth for fatigue detection was RPE (7x), followed by a custom questionnaire, different levels of exercise intensity, and 60% maximal exertion. In these works, subjects had to report their ratings at different intervals, ranging from every 5 *repetitions* (reps) to every 2 minutes. The number of used classes ranges from 2 to 14, while two related works examined multiple classes in the same study. The number of samples (reps)

used for ML ranges from 31 to 2400 with an average of 1273 samples, while 4 of the 10 related works did not report the number of samples.

SVM was utilised most often, followed by RF, LR, CNN, and LSTM, while 6 of the 10 related works examined multiple ML models (see Table B.3 in the Appendix).

Table 2.20 summarises the measures and ML results reported by the related works with squats. 5 out of 10 works used classification, while the remaining 4 works used regression. One of the related works used a CNN model and was evaluated with Pearson's correlation coefficient. *Confusion matrix* (CM) was applied by 3 works.

Tab. 2.20.: Results and measures reported by the related works with squats.

Measure	Acc	Spec	Recall	Prec	F_1	RMSE	MAE	MAPE	R^2	CM
Mean	81.5	-	77.6	76.2	77.0	1.2	1.26	7.9	0.3	-
Median	80.9		77.6	76.2	77.0	1.0	1.26	7.9	0.3	-
Min	76.0		77.6	76.2	71.9	0.6	1.26	7.7	0.2	-
Max	90.0		77.6	76.2	82	2.2	1.26	8.1	0.5	-
Count	6	0	1	1	2	4	1	2	2	3

Table 2.21 shows the applied evaluation types of the 10 related works with squats. The number of input features used for ML ranges from 8 to 122, with an average of 21.8 and median of 12 features. The ratio between training and test sets ranges from 1.6% to 30%, with an average of 18.3% and a median of 15%. None of these works applied *T1-SOLO* evaluation, 3 works *T2-LNSO*, 6 works *T3-LOSO*, and none *T4-LMSO*. Cross-validation was utilised by 9 of the 10 related works. Oversampling was applied by 2 works, while 1 work applied undersampling. Furthermore, 4 related works reported balanced classes, 2 class imbalances, and 4 did not report on class balance.

7 of the 10 related works with squats mentioned small data as a limitation (see Table B.7 in the Appendix). Among them, Albert and Arnrich [9] used a large window overlap of 95% to generate as much training data as possible. Karvekar et al. [189] found that increasing the number of classes in their study led to a decrease in model performance due to overlapping regions within the RPE scale. Gan et al. [126] stated that the ML models need to be assessed in larger sample populations. Jiang et al.

Tab. 2.21.: Number of features, applied evaluation types, cross-validation (CV), oversampling (OS), undersampling (US), and if the classes were balanced (CB) in the related works with squats.

	Features	Test Ratio	*T1-SOLO*	*T2-LNSO*	*T3-LOSO*	*T4-LMSO*	CV	OS	US	CB
Albert and Arnrich [9]	50, 100	6.3%	-	-	x	-	x	x	-	Yes
Gan et al. [126]	18	N/A	N/A	N/A	N/A	N/A	10	-	-	N/A
Perpetuini et al. [275]	9	10.0%	-	-	x	-	x	-	-	N/A
Albert et al. [10]	8	8.3%	-	-	x	-	x	-	-	No
Jiang et al. [176]	32	1.6%	-	-	x	-	x	-	-	N/A
Jiang et al. [175]	10, 6	15%	-	-	x	-	5, 6	Duplicates	-	No
Karvekar et al. [189]	42	N/A	-	x	-	-	5	-	-	N/A
Karvekar et al. [188]	>12	N/A	N/A	N/A	N/A	N/A	-	-	-	N/A
Karg et al. [186]	17	N/A	-	x	x	-	x	-	x	No
Zhang et al. [367]	11	30%	-	x	-	-	5	-	-	Yes
This thesis	23, 67, 122	20%	-	x	x	x	5	SMOTE	-	No

[176] used 12 real subjects and increased the population by additional 50 simulated subjects. 2 related works with squats mentioned data variability: Albert et al. [10] included only male subjects for a homogeneous population. Jiang et al. [176] found no significant difference between fast-tiring and slow-tiring sub-groups. 3 related works with squats mentioned generalisation: Albert and Arnrich [9] and Perpetuini et al. [275] noted that the small data set limits the generalisability of the results. Perpetuini et al. [275] applied nested cross-validation to assess generalisability. Karg et al. [186] stated that the linear model still generalises with unseen data of the subjects despite the small number of 10 to 50 samples per subject.

2.4.4 Summary

A survey was conducted to find 95 primary studies addressing sensor-based fatigue detection with ML during physical activity. The numbers of recruited subjects in the related works ranged from 1 to 80, with an average of 21.1 and a median of 17) subjects. The most common physical activities used in the related works studies were running, manufacturing tasks, walking, and squats. ECG, IMU and sEMG were the most commonly used sensors. The number of collected samples varied widely, with an average of 8759.9 and a median of 1046 samples. RPE scales were used most commonly as ground truth by 46.32% of the related works. Four main evaluation methods were identified in the related works, with *T2-LNSO* (49.5%) and *T3-LOSO* (35.8%) being the most commonly used, while *T1-SOLO* (9.5%) and

T4-LMSO (6.3%) were rarely used. Cross-validation was commonly used with 5 or 10 folds.

83.2% of the related works performed classification and 15.8% regression to evaluate the ML models, with two doing both. The dominant ML approach was supervised. SVM, RF, ANN, *k*-NN, LR, DT, CNN, and LSTM were among the most commonly used ML models. The average accuracy of the ML models in the related works was 85.7%, with a median of 89.2%. Specificity, sensitivity, F_1 score, and confusion matrices were also used, but only in about one third of the related works. The type of F_1 score used (micro, macro, or weighted) was often not explicitly reported. The median test rate was 15%. The number of fatigue classes to be detected ranged from 2 to 14, with binary classification being the most common (57.9%). Imbalanced classes were reported by 24% of the related works. 11% of the related works applied oversampling or undersampling techniques to address imbalanced data.

The topics small data, variability, overfitting, and generalisation were mentioned in 77.77% of the related works, but none of them address these topics in depth.

2.5 Research Gaps

This section highlights the research gaps identified based on the related works. Elshafei et al. [108] found a lack of research on sensor-based fatigue detection, although HAR systems are part of everyday life; despite the abundance of literature in this area, little is known about the impact of muscle fatigue on the performance of these systems. Enoka and Duchateau [110] noted that despite the growing interest in fatigue, little is known about its effects on human performance. This lack of studies on fatigue detection is also highlighted by other works [156, 241, 326, 247]. Enoka and Duchateau [110] observed that fatigue can limit human performance, but there are considerable gaps in knowledge of the underlying mechanisms and

how to manage them. This dilemma appears to be largely due to the inability of current terminology to address the range of conditions attributed to fatigue.

Halson [143] highlighted the lack of generalisable markers of fatigue: "There are a number of potential markers that can be used to gain an understanding of training load and its effect on the athlete. However, very few of these markers have strong scientific evidence to support their use, and no single definitive marker has yet been described in the literature". Closely related to this comment is reproducibility – the ability to verify the generalisability of published findings through independent third-party verification. According to Zheng and Stodden [372], work has been done on technical and cyberinfrastructure solutions for reproducible ML, but research on incentives to adopt reproducible practices lags behind.

The survey in Section 2.4 has shown that *T2-LNSO* and *T3-LOSO* evaluations are commonly used. *T4-LMSO* has only been used in 6 related works, but so far there has been no comparison between all of these evaluation methods. This leads to another research gap on how the evaluation methods affect ML models trained on small data and the resulting generalisability, especially in light of the generally high performance results reported.

Small data sets are mentioned to varying degrees in the related works, often as a limitation. It is often assumed that ML models trained on small data are generalisable as long as they do not overfit, which will be discussed further in this thesis. Vrigkas et al. [339] argued that learning human activities from very little training data or missing data is challenging. Several issues, such as the minimum number of learning examples for modelling the dynamics of each class or safely inferring the activity label performed, are still open and need further investigation. More attention should also be paid to the development of robust methods under the uncertainty of missing data, either on training steps or on testing steps. According to Iwana and Uchida [166], there is a need for more publicly available time series data.

Generalisability and variability are rarely addressed in the related works, and if so, briefly. Hussain et al. [162] noted that current solutions for activity recognition face

the challenge of variability. Variability means that the same activity is performed by a different person or the same activity is performed by the same person at a different pace. Many existing systems cannot deal with the variability problem, i.e., if the same activity is performed by a different person, the system's recognition accuracy is very low. Also, if the same person performs the same activity in a different style, the system's performance degrades. Modern systems should be robust and deal with the issue of variability. This is still an open question and requires further research.

2.6 Summary

This chapter explored the key concepts of this thesis. Small data refers to data sets collected from a small number of subjects and is used for ML to make predictions at the individual level. Fatigue is a multifaceted phenomenon that has been studied in various fields. Despite its widespread importance, there is still no universally accepted definition of fatigue, nor standard methods of objective and subjective measurement. HAR is presented as sensor-based detection of activity, including the state of the individual, such as fatigue. In addition, several taxonomies were introduced for each key concept.

In a survey, 95 related works were found that investigated fatigue detection in physical activity based on sensors and ML. The survey showed that RPE scales were most commonly used as ground truth, as well as binary classification, with supervised ML being the dominant approach. Furthermore, four common evaluation types were identified: *T1-SOLO*, *T2-LNSO*, *T3-LOSO*, and *T4-LMSO*, with *T2-LNSO* and *T3-LOSO* being the most common. The average prediction accuracy in the related works is 85.7%.

In addition, the research gaps were described, including the lack of robust evaluation methods and the need for more attention to generalisation and variability in regard to ML with small data.

3 Fatigue Recognition Chain Framework

According to the Cambridge Dictionary, a framework is a system of rules, ideas, or beliefs used to plan or decide something[1]. This chapter introduces the Fatigue Recognition Chain framework, which is intended as a general guide for interdisciplinary researchers to conduct sensor-based exercise fatigue detection research with small data. The framework covers the entire process from specification, collecting raw data, ML, evaluating generalisable performance of ML models, and sharing of the research. The framework is inspired by the works of Usama Fayyad [333], Bulling et al. [57], Lin et al. [233], Maman et al. [242], Rauschenberger and Baeza-Yates [284], Chicco et al. [75], and Zheng and Stodden [372].

The Fatigue Recognition Chain describes an incremental process [333] of seven steps (see Figure 3.1). If unanticipated obstacles arise, researchers are encouraged to go back to previous steps to make the necessary adjustments before moving on to the next step. Each step contains sub-steps that can be adapted to suit different research needs and workflows. These sub-steps can be skipped, implemented in any order, performed in parallel, or even performed multiple times, providing flexibility to accommodate different research scenarios. Although the Fatigue Recognition Chain provides a flexible structured approach, there may be cases where this framework is not fully applicable. For this reason, it is recommended that preliminary trials are carried out to assess the feasibility of a planned approach and to ensure that it can meet the objectives.

[1] https://dictionary.cambridge.org/dictionary/english/framework

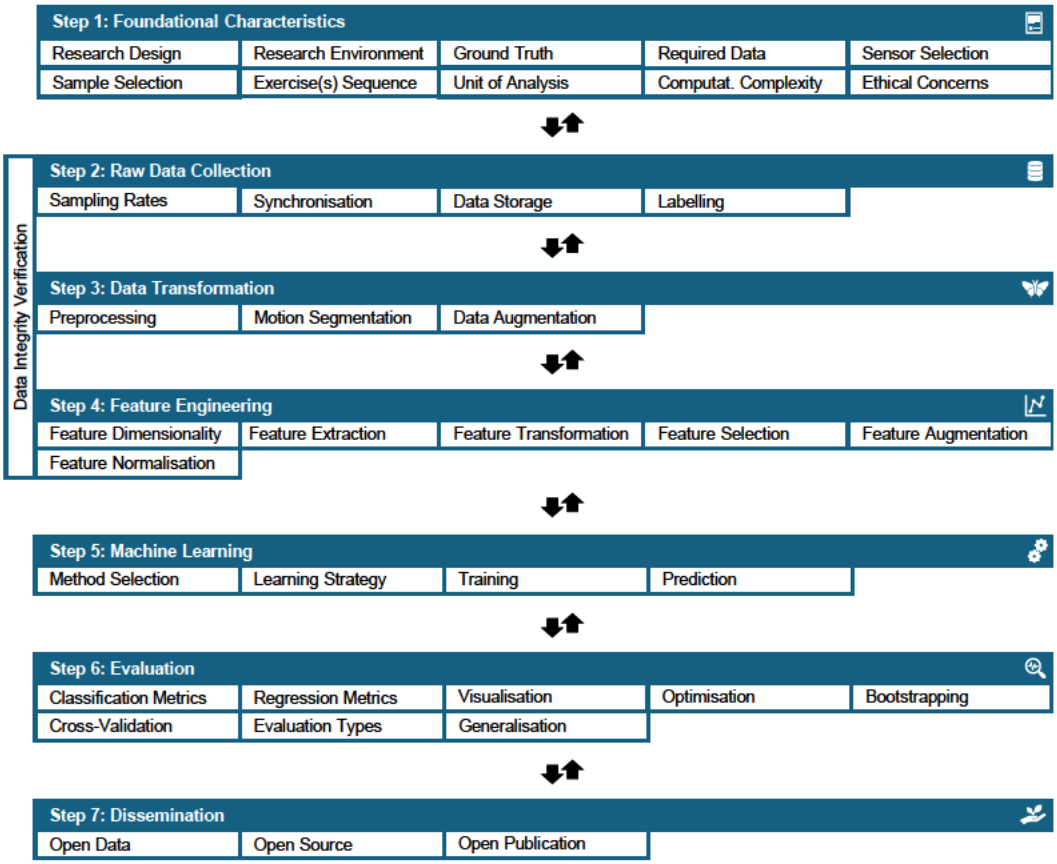


Fig. 3.1.: The Fatigue Recognition Chain.

The Fatigue Recognition Chain can be applied in training and application mode. In training mode, an ML model is trained under prepared conditions. In application mode, the trained model is utilised for evaluation or real-world deployment.

3.1 Step 1: Foundational Characteristics

Step 1 of the Fatigue Recognition Chain addresses the foundational characteristics of the research project, including a clear problem statement and method, which need to be carefully defined upfront as they have a significant impact on subsequent modelling steps [372]. As emphasised by Géron [128], the selection of appropriate performance measures for model training and evaluation is crucial, as is early validation of performance assumptions to avoid misaligned system outcomes.

3.1.1 Research Topic & Design

A well-defined research topic is usually developed iteratively, involving continuous refinement of the research idea until it is fully articulated and accepted. This process involves formulating and clarifying research questions, defining a related aim, and setting specific research objectives. The research design is the general plan of how to answer these research questions (see also the research onion in Figure M.1 in the Appendix) [295, 110].

3.1.2 Research Environment

A research environment should enable continuous monitoring and the generation of reliable responses at the right time to ensure the validity of the results provided. The time at which the experiments are conducted may be important, as it may affect the participants, for example due to temperature, humidity, noise, or circadian rhythms [130, 184, 201, 317]. Furthermore, it is important to consider the natural environment in which the physical activities are performed. For example, performing experiments in a laboratory environment that does not aesthetically resemble a gym may yield different data than performing the same experiments in a real-world setting [130, 339, 255]. Other factors to consider are the ease of use of the sensors, the preparation and conduction effort, logistics, required personnel and equipment, and cost [57].

3.1.3 Ground Truth

Ground truth can refer to a fundamental truth, the real or underlying facts, or information that has been collected at source, however, the origin and interpretation of the term is debatable [352]. In ML, ground truth is usually obtained through mathematical formulae or manual (subjective) labelling. The latter is often time-consuming and labour-intensive; for example video-based frame-by-frame analysis of

training exercises. When labelled data is limited, techniques such as semi-supervised, unsupervised, or transfer learning may be an option [57].

While a clear understanding of the target activities and their characteristics is essential [57], objectively correlating physiological variables with perceived fatigue remains challenging [213]. Despite numerous proposed fatigue detection methods, a universally accepted gold standard has yet to be established [248] (see Section 2.2). Non-invasive methods for fatigue assessment are typically based on the following principles: subjective measures, performance-related methods, bio-mathematical models, behaviour-based methods, and physiological signal-based methods. Each approach has distinct advantages and limitations, with varying suitability as ground truth [248]. Further details of fatigue assessment can be found in Section 2.2.

3.1.4 Required Data

The data needed to answer the research question should be specified in terms of type, structure, unit, quality, quantity, and variability [57, 295, 333]. If new data need to be collected, a systematic set of rules should be defined and followed. Data should be stored in an appropriate format, in an appropriate database, anonymised to protect personal information and, where possible, made publicly available to facilitate further research (requiring the consent of the subjects). How much data a research project needs is a challenging question. So far, there is no general answer or formula.

The use of publicly available databases can reduce the burden of data collection. A variety of data sets are available for HAR, with different sensors, sensor placements, and activities, however, a significant proportion of them were recorded in controlled environments [339, 273]. Another limitation is the availability of labelled data. For example, current activity and fatigue detection data sets do not represent and label every variation of a target activity. This is largely due to the immense variability in human movement, environmental factors, sensor devices, and experimental

setups [162]. Existing data sets may also be limited, inadequate, and imbalanced, posing significant challenges to the development of effective models and analysis. Moreover, the accuracy and generalisability of a fatigue detection model is likely to be inadequate, if the data set does not represent the same data distribution as that of the target application [62]. Another potential source of data could be simulations [176], but such data may not accurately represent real data.

3.1.5 Sensor Selection

Automated fatigue detection requires sensors that can detect physiological or behavioural signals, such as heart rate, muscle activity, or movement patterns. Sensors allow researchers to obtain objective measurements. They can be categorised based on their deployment location: wearable, object-bound, or environment-bound [162]. However, the lack of a standardised definition of fatigue may lead to a potential reality gap due to the indirect nature of fatigue measurement (see Section 2.2). For this reason, multiple sensors are often necessary to capture the full spectrum of fatigue indicators, including a comprehensive range of physiological and motion data relevant to fatigue. Additional sensors may also be required for specific purposes, such as segmentation, labelling, and environmental monitoring [233, 57]. Furthermore, each sensor has different capabilities and possible configurations, resulting in trade-offs between accuracy, system latency, data storage, power consumption, and processing power. In some systems, low-latency classification and immediate feedback are paramount, while in others these factors may be less important. The variability in sensor characteristics, including hardware errors or failures, sensor drift, operating temperature changes, and loose straps, as well as sensitivity to environmental conditions, should be addressed to ensure uninterrupted monitoring and reliable performance [57]. Further details of sensors, commonly used in fatigue research, can be found in Section 2.2.4.

3.1.6 Sample Selection

Sample selection is particularly important in ML with small data, because it controls the variability of the data and can either improve or weaken the generalisability of ML models. Selection criteria in fatigue research are based on subjects and may include factors such as age, gender, fitness level, exercise frequency, exercise experience, and occupational background. Variability in subjects (and data) also affects the number of similar subjects required. Variability affects the required sample size. If a system needs to support a more diverse population or complex exercises, larger data sets are required to capture the full range of variability. User-specific systems can reduce overall data variability by tailoring models to individual users, potentially leading to higher performance for those specific individuals [57].

3.1.7 Exercise(s) and Sequence

The choice of exercises should be based on the specific objectives of the study and the targeted muscle groups to be fatigued. For example, exercises such as squats primarily target the lower body, whereas push-ups primarily target the upper body. Other considerations may include aerobic and anaerobic training. By carefully selecting exercises based on the desired criteria, researchers can increase the likelihood that the data collected will accurately reflect the desired physical activity and fatigue patterns.

The order in which exercises are performed can also affect fatigue assessment. A randomised sequence can minimise bias and account for potential order effects, while a standardised sequence can improve consistency and comparability between subjects. Attention should be given to biomarkers with slow response and decay times, such as heart rate, which is considerably affected by the order in which the exercises are performed. In addition, factors such as exercise intensity, duration, and rest periods between exercises should be standardised or carefully controlled to reduce interfering variables and to improve the reproducibility of results. However,

the exercise protocol should also reflect realistic scenarios that are as close as possible to the targeted application. This often requires a balancing act between controlled and realistic exercise scenarios.

3.1.8 Unit of Analysis

In the context of physical activity, a unit of analysis represents the individual data point or set of data points being studied, such as a repetition, a set, a session, a specific time interval during exercise, or physiological measures. For repetitive exercises, the unit of analysis – or segment – is usually a repetition of an exercise or a set of repetitions. A precise definition of the unit of analysis is essential to obtain consistent and reproducible results. However, the definition is often subjective, dependent on the algorithm, and influenced by the specific application. Once a definition has been established, a specific segmentation algorithm can be designed [233].

3.1.9 Computational Complexity and Storage Requirements

The overall computational complexity of fatigue detection systems is influenced by several factors, including signal processing, data preprocessing, segmentation algorithms, fatigue prediction algorithms, and hardware capabilities. All of these factors can have a cumulative effect on the overall computational effort required. Scalability is a related concern: many algorithms exhibit a disproportionate increase in computational time as the dimensionality of the feature set or the number of templates in the motion library increases [233]. A fatigue detection system can operate in different modes, each with a different impact on computational complexity and memory requirements:

Online systems: Sensor data are collected and processed in real time. These systems are typically optimised for limited computational resources and therefore do not require the full observation sequence before performing segmentation or

fatigue detection, and often operate with limited data windows. Online approaches are often used in interactive applications [57].

Semi-online systems: The ML model is trained offline and then applied in an online context. Some offline algorithms are non-causal and require the entire observation sequence to be available for processing [233]. In addition, model training could be performed continuously in an online manner, with the model being adapted in real time as new data become available.

Offline systems: First, all sensor data are first recorded. Then, fatigue detection is performed. Offline systems are typically used in non-interactive applications such as health analysis [57].

3.1.10 Ethical Concerns & Consent

The implementation of fatigue detection systems raises ethical concerns that need to be addressed to protect individual privacy. One of the concerns is the perceived intrusiveness of these systems, as they often require continuous monitoring of physiological and behavioural data. To minimise intrusiveness, sensors and data collection methods should be as minimal and unobtrusive as possible [330, 57].

The side effects of monitoring devices in private and public spaces should also be considered. Users may feel uncomfortable or anxious about being monitored, which could affect their behaviour and overall well-being. Fatigue detection systems should be designed with respect for personal space and comfort. When used in public settings – such as laboratories, schools, or workplaces – fatigue detection systems need to follow ethical guidelines to avoid misuse or discrimination by ensuring that the system does not have a disproportionate impact on certain groups [330, 377].

The collection, storage, and analysis of sensitive personal data, such as health and activity information, requires additional privacy measures. Such data must be anonymised and stored securely to protect individuals' identities and prevent unauthorised access [330]. In addition, transparent data use policies should be

established to inform users about what data are collected, how the data are used, who has access to the data, and how long the data are stored. Participants should be required to sign a declaration of consent, ensuring they are aware of and agree to these terms. These terms should also be reviewed by an independent ethics committee [348].

Ethical considerations also extend to the transparency and accountability of the deployed detection system. Users should be informed about the algorithms and decision-making processes used to detect fatigue [330].

3.2 Step 2: Raw Data Collection

Step 2 of the Fatigue Recognition Chain addresses raw data collection, another fundamental step, as it directly influences subsequent steps and ultimately determines the performance of an ML system [91]. Raw data are data points that have not undergone any processing and have been collected directly from a specific source, such as a sensor signal or pixels within a video frame. The result of raw data collection is typically an n-dimensional vector, where each row represents a new sensor reading, for example, a timestamp with accelerometer values (see Figure 3.2)[2]. The following vector notation is commonly used to describe sensor output: $s_i = (d^1, d^2, d^3, ..., d^t)$, for $i = 1, ..., k$ where k denotes the number of sensors and d^i denotes the multiple values at time t. Each sensor is sampled at regular intervals, resulting in a multivariate time series [57].

Raw Data Set

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z
0	0.34	0.56	0.13
4	2.79	0.23	0.78
8	0.22	0.01	0.45
16	0.53	0.14	0.75
...	...	...	...

Fig. 3.2.: Example of raw data time series from a 3-axis IMU. Each row is a vector. Each column is a variable/dimension. Each cell contains a data point.

[2]The representation of time series in column vectors is purely conventional.

3.2.1 Sampling Rates

Raw data is typically collected as time series data at a specified sampling rate. Adopting a high sampling rate may provide sufficient information for detailed data analysis, but it also burdens the system with large data sizes and increased computational load. Conversely, using a low sampling rate may not capture the intrinsic characteristics and may miss important nuances. There is no consensus on the optimal sampling rate for motion-sensitive sensors, as it often depends on the specific application and the nature of the activity being monitored [79]. Table B.1 in the Appendix shows the various sampling rates for different sensors applied by related works.

3.2.2 Synchronisation

Synchronisation of data signals from multiple sensors is essential [57]. Maintaining synchronised clocks across all sensor units, which can drift over time, is a special area within distributed systems in computer science [78, 224, 219]. The accuracy required for clock synchronisation varies depending on the application. Some scenarios require strict accuracy, while others allow more flexibility. In addition, synchronisation may involve different sampling rates, which may require upsampling, downsampling, interpolation, and/or wavelet transformation techniques.

3.2.3 Data Storage

There are specialised databases for querying, storing, sharing, managing, and analysing different types of data, such as InfluxDB or MongoDB, which support time series data[3].

[3] https://en.wikipedia.org/wiki/Time_series_database

3.2.4 Labelling

Labels are high-level attributes assigned to data points that represent quantities of interest, such as fatigue levels [254]. The term 'label' is often used interchangeably with 'class', 'response', 'output', or 'target variable'. The acquisition of labelled data can be a costly process, requiring the input of human experts [179].

The majority of related works applied supervised ML (see Section 2.4.2), which requires a labelling process. Data labelling involves establishing guidelines, categorising classes, labelling tools, and storage pipelines [62]. Labels are often added as an additional dimension to a data vector, typically at the end by convention (see example in Figure 3.3)[4]. Labels can be created during or after data collection, either

Labelled Data

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z	Label
0	0.34	0.56	0.13	8
4	0.28	0.23	0.78	8
8	0.22	0.01	0.45	8
12	0.42	0.08	0.60	8
16	0.53	0.14	0.75	10
...	...	...	...	

Fig. 3.3.: Example of labelled data in a time series.

manually by humans or automatically by machines (see also Section 2.2.4). Manual labelling is limited by the frequency and complexity of the labels, with more complex labels requiring more analysis time. Domain experts can be recruited to label the data either ad hoc or post hoc, for example through video or graph analysis. This process may also be outsourced to crowdsourcing services, but validity can be an issue with such services. Another source of labels is the subjects, who typically provide labels by voice or questionnaire, although this method also has limitations in terms of frequency [179, 128]. Automatic labelling offers advantages such as increased consistency and frequency, support for real time systems, and minimal additional effort. However, not all labels can be assessed automatically, particularly subjective and qualitative labels, which are difficult to encode into an algorithm.

[4] Label information may also be stored separately with a start and end timestamp to reduce redundant data and thus the amount of storage required.

A label is typically denoted by y (if it is a single number) or $\mathbf{y}$ (if it is a vector of different label values). Numerical labels are used for regression, while categorical labels are used for classification. In binary classification, each data point or sample belongs to exactly one of two classes. In multi-class classification, data points or samples belong to exactly one of more than two categories. In addition, there are applications where data points can belong to several categories simultaneously. Ordinal labels fall between numeric and categorical labels, where labels are represented by sequential numeric values. For example, 0 might indicate no fatigue, 1 some fatigue, and 2 exhaustion [179].

One challenge is that fatigue is a continuous process and the exact boundaries defining its start and end are often ambiguous. To overcome this problem, different labelling methods are used, as shown in the literature review (see Section 2.4.2). Some related works used automated labelling approaches (e.g., thresholds) and some relied on manual labelling approaches (e.g., RPE). In unsupervised learning, the data is typically not explicitly labelled, as there are no fixed targets given to the ML model during training. Instead, the model identifies patterns and structures in the data without relying on external labels. However, certain tasks, such as clustering, can be thought of as implicit labelling.

3.2.5 Data Integrity Verification

While ML automates data analysis, the quality of the output depends on the quality of the input data. The phrases "garbage in, garbage out" and “data $\neq$ information” underline the fact that inaccurate, biased or incomplete data used for ML is likely to lead to unreliable predictions [25, 75, 148, 91]. Data integrity ensures that data is accurate, consistent, and reliable throughout its lifecycle. This includes verifying that the assumptions made about the data are valid, and that the data have not been tampered with or corrupted. Hardware or software failures, human error, and malicious actors can all pose threats to data integrity [180]. For example, time series

data can be affected by various data quality issues due to frequent equipment and transmission failures [318]. In addition to technical issues, distributional shifts and selection bias can be a problem [25].

High quality data is characterised by attributes such as accurate, clean, compatible, complete, easy to access, interpretable, reliable, secure, timely, traceable, trustworthy, unbiased, useful, and valuable. Quality data prevents error propagation and improves model performance and convergence rates. Conversely, outliers, missing values, high dimensionality, varying scales, bias, and privacy concerns can degrade data quality [322]. For statistical reasons, these issues may have a greater impact on small data than on big data.

One approach for achieving high quality data, often feasible with small data, is initial data exploration by visually inspecting the plotted data for anomalies [75]. Exploratory Data Analysis is a related approach, originally from statistics, used to understand data and prepare it for further analysis. It involves initial exploration to discover patterns, identify anomalies, and test hypotheses using statistical techniques and visualisations and can help to ensure that the data is well prepared for subsequent ML methods [90, 197, 268].

However, these approaches are usually manual and time-consuming. As a result, only a limited number of samples can be validated in a reasonable amount of time. To overcome these limitations, data constraints can be defined to automatically validate data against specific (statistical) criteria, such as plausible value ranges (e.g., joint angles). Constraint-based validation can increase data reliability and avoid erroneous results [318, 75]. However, the constraints depend on the particular application and the desired data quality standard [322]. Analogous to unit testing in software development [196], data integrity verification should be performed regularly and, where possible, automatically. For this reason, data integrity verification is a sub-step that spans several steps (2–4) in the Fatigue Recognition Chain.

3.3 Step 3: Data Transformation

Step 3 of the Fatigue Recognition Chain addresses data transformation. According to Kahneman [182], data is incomplete, dirty, and noisy, and it takes most of the time to curate it. The overall goal of the data transformation step is to refine the collected raw data into a form suitable for effective feature extraction and subsequent ML [57, 233]. If a data set contains a considerable number of errors, outliers, or noise, it will prove more challenging for an ML method to identify the underlying patterns. Consequently, most data scientists spend a significant amount of time on data cleansing and transformation [128]. The following sub-steps aim to enhance the informative characteristics of a data set. These sub-steps are often interdependent, which means that their order of execution can be important.

3.3.1 Preprocessing

Preprocessing[5] involves the transformation of data into a more structured and usable format. This process may include various techniques such as cleaning, filtering, calibration, normalisation, resampling, synchronisation, segmentation, and signal-level fusion [322, 75, 57]. For example, preprocessing may find and replace missing data or remove noise, outliers, and artefacts while preserving the essential information [75]. Artefacts can come from a variety of sources, such as physical activity disturbances or sensor malfunctions. The reduction of noise is often required due to sensor variability and limited digitisation processes [130, 350]. For specific sensors such as accelerometers, specific signal processing techniques such as noise reduction and baseline drift correction can be applied [57]. According to Kokol et al. [202], the most commonly employed preprocessing techniques to overcome small data are linear and non-linear principal component analysis, discriminant analysis, data augmentation, virtual sample, feature extraction, and autoencoder. A comprehensive survey of preprocessing techniques for time series can be found in

[5] Preprocessing and signal processing are two related concepts.

Tawakuli et al. [322]. Figure 3.4 shows an example of raw data transformed into a vector series, with missing data interpolated and outliers corrected.

Raw Data Set

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z
0	0.34	0.56	0.13
4	2.79	0.23	0.78
8	0.22	0.01	0.45
16	0.53	0.14	0.75
…	…	…	…

Outliers and missing data interpolated

Preprocessed Data Set

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z
0	0.34	0.56	0.13
4	0.28	0.23	0.78
8	0.22	0.01	0.45
12	0.42	0.08	0.60
16	0.53	0.14	0.75
…	…	…	…

Fig. 3.4.: Example of preprocessing, with missing data interpolated and outliers corrected.

The goal of preprocessing is to prepare the data for analysis while ensuring that essential signal characteristics, which convey critical information about the activities of interest, are preserved. In addition, preprocessing can be used to support other tasks. For example, a signal filter can be used to facilitate a segmentation algorithm. A model could then be trained on the raw data using the identified segments. In general, preprocessing should be generic, i.e., it should not depend on anything other than the data itself, e.g., it should not be specific to a particular person [57]. On the other hand, overly aggressive data correction can remove relevant nuances in the data set, which can corrupt existing patterns and negatively affect the outcome of ML.

Filtering

Sensor data often contain noise due to miscalibration, malfunction, placement errors, environmental conditions, or multiple activities. Common noise reduction techniques include low-pass, mean, linear, wavelet, and Kalman filters [91]. *Finite Impulse Response* (FIR) and *Infinite Impulse Response* (IIR) filters are two basic types

of digital filters typically used in signal processing. The main purpose of FIR and IIR filters is to manipulate or modify a signal by removing unwanted components (e.g., low-/high-/band-pass filters), enhancing certain characteristics, or preparing the signal for further processing [76, 234, 266]. Both types of filters have specific advantages depending on the application. FIR filters (e.g., moving average) are known for their inherent stability as they have no feedback loops. In addition, FIR filters can be designed to have a linear phase response, ensuring that all frequency components of the signal are delayed by the same amount of time, preserving the signal's waveform shape. However, FIR filters often require more computing resources than IIR filters (e.g., Butterworth) because they can involve a greater number of coefficients to achieve the desired frequency response. Moreover, the design process for FIR filters is generally easier in many applications (see also Appendix L) [234, 266].

Interpolation

Timestamps in a data set can vary due to data corruption, hardware problems or connection errors. These discrepancies can lead to misaligned data streams, which can affect the reliability of algorithms (e.g. Butterworth filters). Interpolation is a mathematical technique used to estimate unknown values within the range of known data points. Interpolation techniques, such as linear interpolation, can be applied to infer missing or misaligned data points. Interpolation uses known data points to construct new data points within the range of a discrete set of known data points, effectively filling in the gaps caused by irregular timestamps. More sophisticated interpolation methods, such as polynomial or spline, may be used, depending on available computing resources and the complexity of the data [340].

Sample Rate Conversion

Sample rate conversion is a common technique used to align the sample rates of different data sources. Downsampling reduces the sample rate by dropping some data points, which can also help manage storage requirements when dealing with high frequency data. Upsampling, on the other hand, increases the sample rate by adding interpolated data points, ensuring that lower frequency data can be aligned with higher frequency signals. This alignment is essential when integrating data from multiple sensors to keep all data streams comparable [277]. Another approach is the transformation of time series into a spectral or wavelet representation.

Unit Conversion

When dealing with data from multiple sources, different units of measurement can introduce inconsistencies and make analysis more difficult. Some raw sensor signals can also be difficult to interpret. For example, raw IMU signals are typically converted to acceleration in m/s^2.

Data Fusion

Multimodal data refers to information collected from different sources or modalities, including text, images, audio, numeric, biometric and behavioural data. Each modality can provide unique and complementary insights. By integrating data from different modalities through data fusion, it is possible to enhance model performance and gain a more comprehensive understanding of the subject or phenomenon under investigation Pawłowski et al. [270] and Bian et al. [39].

Data fusion algorithms are commonly classified according to the level of information abstraction: signal level, feature level, or decision level. At the signal level, multiple signals are combined to produce one or more signals of the same form but with better quality. Alternatively, fusion can be performed after feature extraction. Decision-level fusion represents the highest level of abstraction and is often used

when the signals are dissimilar [177, 104]. However, a universal approach to data fusion has not yet been established. According to Pawłowski et al. [270], all data fusion models suffer from the following main problems: they are either task-specific or overly complicated, and they often lack interpretability and flexibility.

Windowing

Sensor data can be processed in windows to reduce computational complexity and data storage requirements. In this context, windows refer to subsets of data collected over a period of time or of a defined size. This approach is useful in resource-constrained or real time scenarios where limited processing power and memory require efficient data processing. The two main variables in windowing are the size of the window and the amount of overlap between adjacent windows. The window size can be either fixed or dynamic, allowing flexibility depending on the application and desired level of detail [233]. For small data sets, windowing may be less critical due to the smaller sample size and thus reduced computational and storage requirements, although this depends on the specific characteristics of the data set. On the other hand, a large window overlap can also be utilised as a kind of bootstrap to generate more training data Albert and Arnrich [9].

Class Reduction

To improve ML model performance with small data, class reduction is often used [212]. This involves grouping similar classes into broader categories, allowing the ML model to focus on more distinctive patterns. However, overly broad categories can obscure important distinctions within the data, leading to a loss of information. It's also important to preserve the original distribution of the data, otherwise the transformed data may no longer be representative.

3.3.2 Motion Segmentation

Segments group individual data points and describe contextual information to distinguish individual movements [91]. Motion segmentation can be useful for detecting exercise fatigue by providing insight into changes in motion patterns over time, which may be indicators of fatigue onset and progression [247, 189, 26, 175]. It is probably for this reason that many related works rely on motion segmentation, where a sample for ML often represents an exercise repetition (see Section 2.4). Motion segmentation can be applied to a variety of sources, including video playback, time series, secondary proxy sensors, or event-based data [57]. It involves identifying the temporal boundaries of motions of interest, decomposing a continuous sequence of motion data into smaller components, and determining the start and end points of each motion primitive (i.e., segment or repetition). Furthermore, motion segmentation allows to distinguish between repeated motions and transitions between different types of motion. Segments can also be labelled with the appropriate motion type and/or quality [233, 91]. Figure 3.5) shows an example of a segmented time series, where a dimension is added to associate data points with segments[6].

Segmented and Labelled Data

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z	Segment ID	Label
0	0.34	0.56	0.13	1	8
4	0.28	0.23	0.78	1	8
8	0.22	0.01	0.45	1	8
12	0.42	0.08	0.60	1	8
16	0.53	0.14	0.75	2	10
…	…	…	…	…	

Fig. 3.5.: Example of segmented time series.

Segmentation Generalisability

A segment should be defined in a manner that facilitates the creation of a manual or automatic algorithm capable of generalisable and consistent (reproducible)

[6] Segment information may be stored separately, including start and end timestamps, to reduce data redundancy and storage requirements.

segmentation of the collected data. Segmentation generalisability refers to the ability of an algorithm to perform on data that differs from the training set [233]. The generalisability of segmentation methods can be divided into intra-subject, inter-subject, and inter-primitive variability. Intra-subject segmentation refers to consistent segmentation for the same subject. Inter-subject segmentation is able to reliably segment across multiple subjects. Inter-primitive variability occurs when the training data is obtained from one set of motion primitives, while the test set consists of a second set of unseen primitives [233].

Segmentation Categories

A segmentation method consists of four components: filtering and outlier rejection, feature space transformation, segmentation mechanism, and identification mechanism [233]. While various segmentation approaches exist, this section outlines common motion segmentation categories.

By Technique Segmentation techniques can be window-, energy-, or proxy-based. In a sliding window approach, a window is moved sequentially over the time series to extract segments. The choice of window size, window step size, and overlap affects segmentation accuracy and computational efficiency [57, 91]. Dynamically adjusting the window size can improve performance, particularly in scenarios where motion patterns change over time, such as during fatigue onset [108]. Energy-based segmentation leverages the varying energy levels within sensor signals to identify different activity intensities and segment the data accordingly by applying thresholds to the energy levels, for example, by having a subject perform a predefined rest position between activities [57, 91]. However, these thresholds may need to change over time, as average muscular endurance typically decreases due to fatigue [108]. Proxy sources refers to additional sensors and contextual sources to support segmentation [57]. For example, a camera-based tracking system could compute

the absolute joint coordinates during an activity to segment accelerometer signals from a smartwatch.

By Boundary Detection Physical changes occur when the motion in question begins or ends. These changes may include alterations in joint motion direction, changes in contact, or the act of picking up an object. Deciding which joint motion to track or how to handle multiple joint changes introduces an additional layer of complexity, as it is not always obvious which joint to use for segmentation. Derived metric boundaries can be defined by changes in a metric or derived signal, e.g, changes in variance, metric thresholds, or state transitions. Such metrics can be determined by either unsupervised or supervised algorithms. Unsupervised approaches have the advantage of reducing the need for manual data labelling. However, for some segments it may be difficult to derive appropriate metrics. Template boundaries are based on user-provided templates, which are calculated by algorithms such as template matching, dynamic time warping, or classifiers. The creation of a template requires the collection of prior data. Selecting the most appropriate template can be challenging, as it should ideally represent the generalisable motion pattern rather than a marginal case [233].

By Labelling Approach Supervised approaches use labelled data to identify the key features of segments. This approach may be feasible depending on the resources available, such as time, personnel, budget, the volume of data to be labelled, and the effort required to label. Unsupervised approaches do not use labelled data or pre-trained models. Adaptive approaches update the model online as new data is collected [233].

By Processing Requirements Online methods process data in real time without requiring prior training. They often employ simple techniques like thresholding or segment point modelling. Semi-online methods combine offline training with online segmentation, allowing for more complex segmentation models while maintaining real time capabilities. Offline methods require the entire data set for both

training and segmentation, enabling the use of computationally intensive algorithms [233]. Figure 3.6 by Lin et al. [233] visualises the relationship between these three segmentation approaches in terms of training and testing.

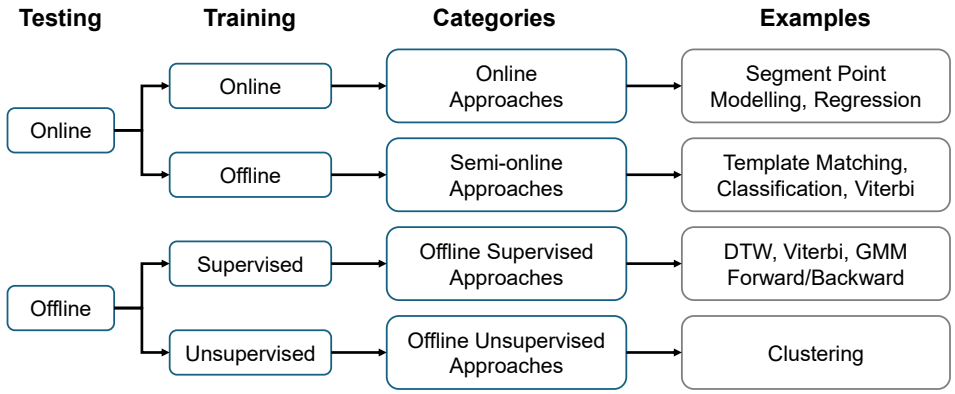


Fig. 3.6.: Overview of the segmentation mechanics by Lin et al. [233].

Segmentation Verification

According to Lin et al. [233], the majority of studies do not perform or report segmentation verification, making comparisons between methods difficult. However, verification of segmentation is necessary to evaluate its performance on a given data set against a ground truth. Ground truth data typically consist of manual segment points labelled by experts. These segment labels are compared with those generated by the segmentation algorithm to calculate metrics such as false positives, false negatives, true positives, and true negatives. Additional evaluation metrics, such as shape similarity between templates and observations, may also be applied. Another aspect is how the segments are labelled. Time series can be labelled either with a temporal tolerance at the segment boundaries or with all data points within a segment. The latter method is less stringent than the temporal tolerance approach, as incorrect segment boundaries do not greatly affect the results, since there are usually many more data points within the segment to smooth out poor segment boundaries [233]. Figure 3.7 illustrates this difference.

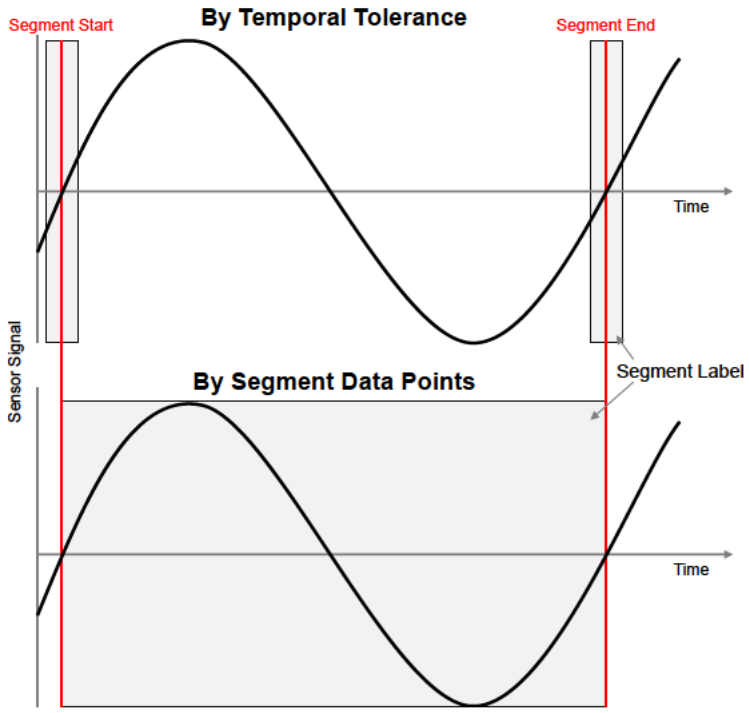


Fig. 3.7.: Segment labelling strategies by temporal tolerance or segment data points.

Segmentation Challenges

Accurate motion segmentation in time series is a challenging task for several reasons. Firstly, human motion has a large number of degrees of freedom. Secondly, human movement is inherently variable. Individual differences in kinematics and dynamics, as well as intra-subject variations due to factors like fatigue or recovery, contribute to this complexity. Overlapping movements, where one action begins before the previous one concludes, further complicate the segmentation process. Additionally, the presence of irrelevant movement patterns within the data can interfere with the identification of target activities. These factors introduce both spatial and temporal variability [57, 233]. Thirdly, generalisability remains a major challenge in motion segmentation. While techniques exist to create multi-subject templates that account for individual variation, their effectiveness is often limited. Applying segmentation algorithms to diverse populations, including individuals with different demographics, abilities, or performing unfamiliar movements, further exacerbates the generalisability problem. This is particularly problematic when training data is

scarce, as most algorithms are evaluated on small data sets, hindering their ability to generalise across subjects and movement types. Fourthly, there is a lack of reported segmentation accuracy and public segmentation databases that would provide a common basis for comparing different algorithms. Such databases could also reduce the amount of post-processing required by researchers to verify algorithms. Finally, some segmentation algorithms require labelled data for training, which can be time-consuming to generate [233].

3.3.3 Data Augmentation

Data augmentation creates new, synthetic data points based on existing ones, increasing data diversity without altering class distribution. Data augmentation is commonly used in fields such as image processing and deep learning. Data augmentation artificially extends data sets through random transformations. There are different augmentation techniques depending on the type of data. Common techniques for images include rotations, flips, translations, scaling, adding noise, cropping, and colour adjustments [210, 309]. For time series, techniques include adding random noise (jittering), inverting the signal (flipping), changing the amplitude of the signal (scaling), distorting the time axis (time warping), extracting random windows from the series (window slicing), randomly shuffling parts of the series (permutation), and applying random distortions to the amplitude (magnitude warping). The suitability of these methods depends on the specific characteristics of the time series data. For example, adding noise assumes that it is normal for the time series patterns of the particular data set to be noisy. A comprehensive taxonomy of time series augmentation can be found in Iwana and Uchida [166] (see also Appendix Q).

3.4 Step 4: Feature Engineering

Step 4 of the Fatigue Recognition Chain addresses feature engineering. Most ML methods require the data to be arranged in a particular representation prior to the learning phase [75]. Feature engineering is the process of transforming data into features with the aim of improving ML and predictions of ML models [208].

3.4.1 Feature Dimensionality

This section briefly introduces the concept of features. It then explores the challenges associated with feature dimensionality.

Features

Features (also known as attributes) are low-level properties of data points that can be measured or automatically calculated. Synonyms for the term feature are covariate, explanatory variable, independent variable, input (variable), predictor (variable), or regressor [179]. The choice of which characteristics to use as features is a design decision [179]. In individual-based ML, features should be robust across subjects and intra-subject variation to be effective [57]. Traditionally, researchers have relied on domain expertise to manually craft features, a process often hampered by subjectivity and limited generalisability [39]. In contrast, deep learning methods can automatically extract relevant features that can outperform hand-crafted features. However, deep learning models can be computationally intensive and require large amounts of data [91].

Balancing feature complexity with computational efficiency is another aspect, especially in real time applications. Minimising the number of features while still achieving the desired performance is a central aspect of feature engineering [207, 208]. However, fatigue can be a challenge in this regard, as fatigue can cause subjects to change their movement patterns over time, which can affect the relevance

of features by reducing their correlation coefficients, rendering some previously significant features insignificant.

Multimodal Features

Combining data from different modalities to create features can improve ML, especially when the data are complementary. The optimal fusion strategy depends on factors such as the nature of the data, computational resources, and the specific application. Early fusion combines data at the feature level, creating a single feature vector. Late fusion integrates information at the decision level, combining the outputs of independently trained models. Slow fusion combines data at intermediate processing stages, balancing the benefits of early and late fusion [339].

Feature Quantity

Any additional feature can introduce noise due to measurement or modelling errors, which can affect the accuracy of the ML method [179]. For example, a data set can contain redundancy because some data are highly correlated [192]. While there are no definitive guidelines on the maximum number of features, a general rule of thumb suggests a much larger sample size compared to the number of features to achieve robust ML. The informal condition $\frac{numSamples}{numFeatures} \gg 1$ can be satisfied either by collecting a sufficiently large number of samples or by using a sufficiently small number of features [179].

Curse of Dimensionality

Bellman [35] coined the term "curse of dimensionality" to describe several phenomena associated with high-dimensional data. A key issue is that as the dimensionality of the feature vector (or feature space) increases, the amount of training data required to estimate model parameters increases exponentially [57, 89]. The variation in the distance between arbitrary points decreases with the addition of more dimen-

sions; consequently, as more features are used to describe the data, data points tend to appear more similar to each other [89]. For most ML methods, the computational cost also increases with the number of features, as does the cost of collecting the data. Therefore, reducing the number of features is beneficial for both theoretical and practical reasons [89].

Dimensionality Reduction

Dimensionality reduction is a technique used to transform high-dimensional data into a lower-dimensional space while preserving essential information. By reducing the number of features, it simplifies the data set, improves computational efficiency, and mitigates the challenges associated with the curse of dimensionality [24]. For example, a data set with n dimensions has 2^n possible feature subsets [192].

3.4.2 Feature Extraction

Feature extraction transforms the raw or preprocessed data set into a new set of features that capture the most important information in a more compact way [57]. This process can also involve mapping complex data structures, such as images or graphs, into a set of features that are needed for some ML methods [75]. Figure 3.8 shows how segments from an accelerometer are transformed into feature vectors.

Feature extraction methods create new features through mathematical transformations or combining existing features, with the aim of improving the performance of ML models [24]. Various types of features can be incorporated, including signal-based features (e.g., statistical [140], frequency-domain [19], wavelet-domain [91], and dynamic features [38]), body model features, event-based features, and multi-level features (e.g., clustering followed by event statistic reduction) [57, 233]. A potential drawback of feature extraction methods is the possible loss of relevant information [192].

Raw Data or **Preprocessed Data**

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z
0	0.34	0.56	0.13
4	0.79	0.23	0.78
8	0.22	0.01	0.45
…	…	…	…

Manually or automatically segmented and labelled

Segmented and **Labelled Data**

Timestamp (ms)	Acceleration X	Acceleration Y	Acceleration Z	Segment ID	Label
0	0.34	0.56	0.13	1	8
4	0.79	0.23	0.78	1	8
8	0.22	0.01	0.45	1	8
12	0.84	0.81	0.33	2	10
16	0.99	0.73	0.41	2	10
20	0.33	0.58	0.72	3	11
…	…	…	…	…	…

Compute features

Feature Set

	Feature A	Feature B	Feature …	Feature N	Segment ID	Label
Sample 1	0.45	0.27	…	0.43	1	8
Sample 2	0.92	0.77	…	0.64	2	10
Sample 3	0.33	0.58	…	0.41	3	11
Sample N	…	…	…	…	…	…

Fig. 3.8.: Example of transforming raw time series into a feature set. Each row is a feature vector (also: sample or observation). Each column is a feature (dimension).

There are three extraction approaches, which can also be combined. (1) Domain-specific approaches rely on expert knowledge to create new features that are predictive for the specific problem, but may be limited by the availability of such knowledge. (2) Explicit or implicit feature mapping transforms data into a higher dimensional space to build non-linear models. For sequential data such as time series, explicit feature mapping can be done in several ways, such as splitting the sequence into sub-sequences (sequential or not, and overlapping or not) and then computing properties of these sub-sequences (e.g., mean, mode, and variance). However, explicit feature mapping can suffer from the curse of dimensionality with excessive (irrelevant) feature generation, which is why implicit methods (e.g., kernel methods) are more popular although less interpretable. (3) Learned feature mapping generates tailored features, e.g., by using deep neural networks, but often requires large data sets [75].

3.4.3 Feature Transformation

Feature transformation is the process of modifying or combining existing features to create new features that better represent the underlying patterns in the data. The aim is to transform the data into a different space where the relationships between the features and the target variable are more apparent. This can be achieved through various techniques such as normalisation, merging features through mathematical operations, or mapping skewed data to a target distribution [253, 376].

3.4.4 Feature Selection

Feature selection and feature transformation are distinct but complementary techniques for reducing the complexity of high-dimensional data. Feature selection identifies and retains a subset of the original features, while feature transformation converts the data into a lower-dimensional space [207]. Figure 3.9 presents a taxonomy of common dimensionality reduction methods, drawing on the works of Cheung and Jia [74], Kathirgamanathan and Cunningham [192], and Pudjihartono et al. [281]. Each of these methods has specific limitations to consider. For exam-

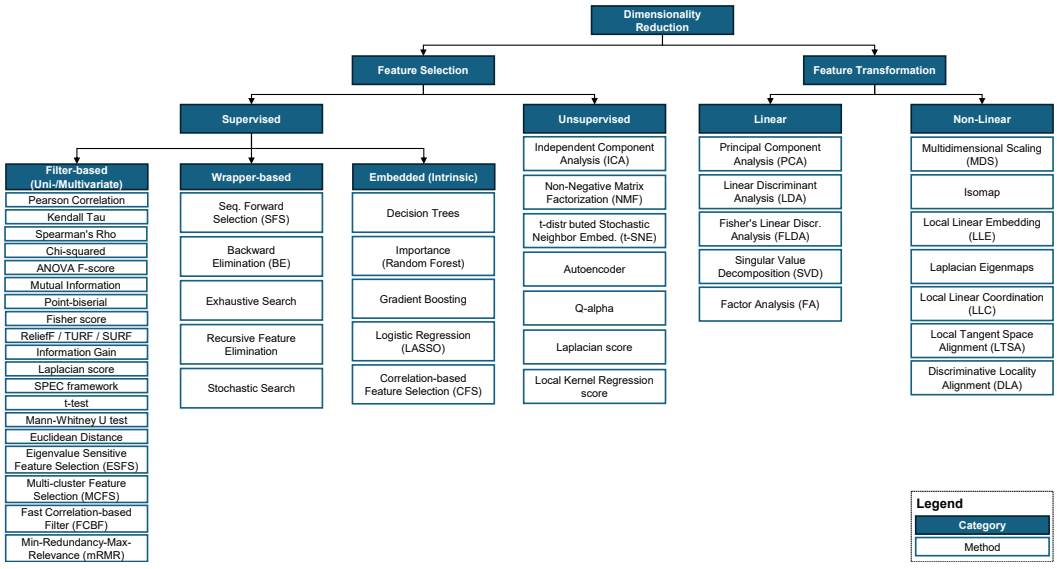


Fig. 3.9.: Taxonomy of common dimensionality reduction methods.

ple, principal component analysis works best with data that follow a multivariate Gaussian distribution, and using the correlation coefficient will only reveal linear relationships, while non-linear correlations may still be present [75].

Feature selection methods can be broadly categorised into supervised and unsupervised approaches [74, 310]. Ang et al. [17] proposed semi-supervised and semi-unsupervised feature selection as extensions to traditional supervised and unsupervised methods. These approaches leverage both labelled and unlabelled data to identify informative features when labelled data is scarce.

In practice, feature selection involves subset selection and feature evaluation. Subset selection aims to optimise subsets of features by removing irrelevant or redundant features [24, 207]. Several studies have shown that there is no universal method for feature subset selection; the optimal approach is often context-specific and requires tailor-made solutions for different problem situations [281]. Figure 3.10 shows an example of subset selection. In contrast, feature evaluation assesses the

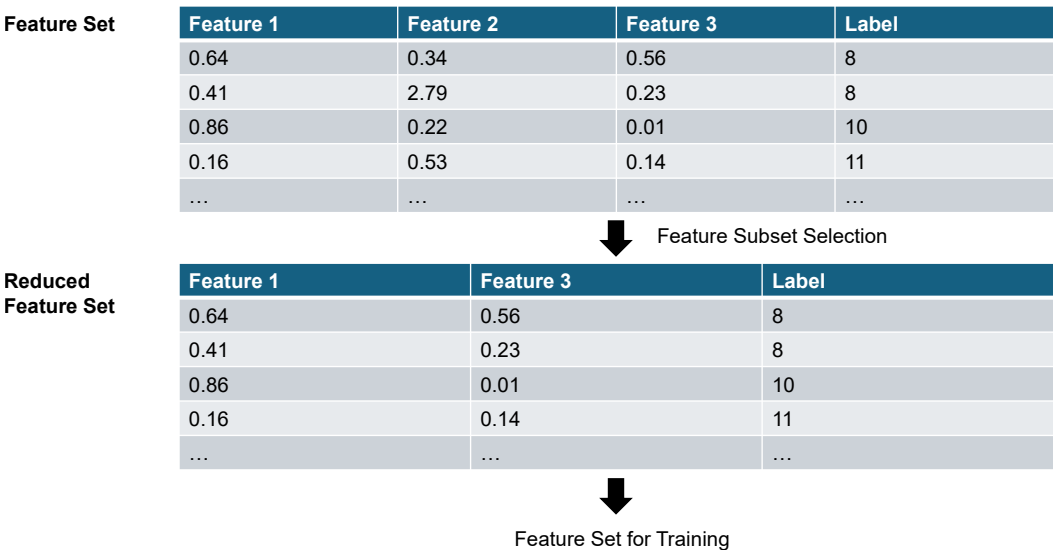


Fig. 3.10.: Example of dimensionality reduction through feature subset selection.

individual quality and relevance of features [287]. Ultimately, both approaches help to improve model performance by focusing on the most informative aspects of the data [17, 207, 208].

Supervised Feature Selection

In supervised approaches for feature selection, the relevance of each feature is assessed based on its correlation with the target variable (i.e., class or label). As illustrated in Figure 3.11, supervised feature selection methods can be divided into filter, wrapper, and embedded methods [74, 287, 281].

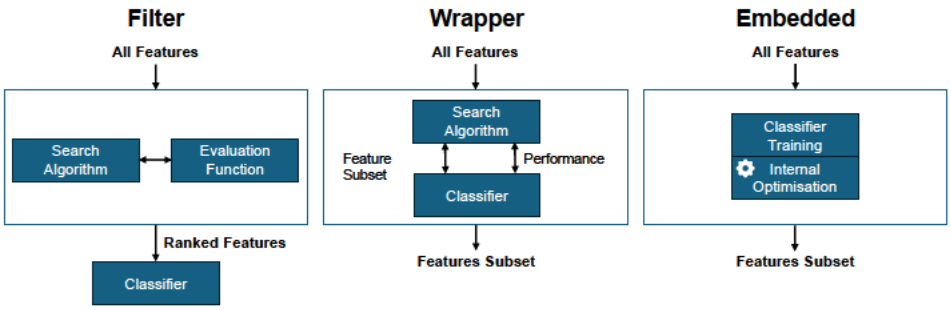


Fig. 3.11.: Supervised filter, wrapper, and embedded feature selection [281].

Filter Filter methods assess feature relevance independently of the ML model by employing statistical measures to rank features based on their correlation or dependence on the target variable [207, 89, 281]. These methods typically assign scores to features and select those with the highest scores based on a threshold. Filter methods can be divided into univariate and multivariate approaches. Univariate methods assess features independently, focusing on individual feature relevance to the target variable. While computationally efficient, they often overlook feature interactions and redundancies. In contrast, multivariate methods consider the relationships between features, potentially identifying more informative and compact feature subsets [207, 281].

Wrapper Wrapper methods iteratively build and evaluate ML models using different feature subsets to identify the optimal combination [207, 91]. While an exhaustive search over all possible feature combinations directly optimises model performance, it is computationally expensive or even impractical and prone to overfitting [89]. In addition, the selected features are specific to the chosen model, limiting their generalisability [281].

Embedded Embedded (or intrinsic) methods integrate feature selection directly into the training process of ML. Methods such as AdaBoost, Lasso, decision trees, and random forests have built-in feature selection mechanisms – they inherently select features that contribute most to improving model accuracy [57, 207].

Hybrid Hybrid methods combine filter and wrapper approaches to leverage their respective strengths [281]. These methods aim to balance computational efficiency and predictive performance by integrating techniques like filter ranking with wrapper optimisation or employing ensemble strategies, i.e., aggregating the results of multiple feature selection methods [192, 281].

Unsupervised Feature Selection

Unsupervised feature selection identifies relevant features without relying on labelled data. By analysing data structure and patterns, these methods aim to select features that capture the inherent variability and relationships within the data [74].

3.4.5 Feature Augmentation

Feature augmentation artificially expands the feature set by applying various transformations to the data. This technique can improve model robustness and generalisation by exposing ML models to a wider range of data variations. This is particularly beneficial for small data sets or data sets with imbalanced classes [166].

Imbalanced Classes

Imbalanced classes refer to a skewed distribution of classes (or labels) in a data set, which can lead to unreliable performance of trained ML models [315, 321]. To accurately assess model performance, normalised confusion matrices or metrics such as precision, recall and F_1 score are essential [57]. Accuracy is not an appropriate metric when imbalanced classes exist [204]. According to Spelmen and Porkodi [315], the causes of imbalanced data are the lack of density in the training data set,

the presence of small disjuncts, the overlapping between classes, the identification of noisy data, the importance of borderline instances, and the data shift between the training and the test distributions. Data sets with biases are likely in health data; gender or age biases are even normal [284]. In the context of exercise fatigue research, data sets are prone to bias because less data is usually collected for the fatigued state [243] – subjects cannot exercise indefinitely in a fatigued state. On the other hand, exercising to exhaustion guarantees the fatigued state for each subject. An alternative is to perform exercises with a fixed amount or time limit, which may result in some subjects not reaching the fatigued state due to differences in fitness and energy levels [4] (see also Fatigue Exercise Load in Appendix D).

Feature Augmentation Approaches

Methods to address imbalanced data include certain loss functions such as focal loss [233], class weights to reduce the influence of majority classes [375], or balanced accuracy that highlights the low performance of minority classes [54]. In addition, the experimental procedure could be designed in such a way that equal class distributions are more likely or even guaranteed. Another approach might be to reduce the total number of classes [57].

Imbalances can also be reduced by undersampling the majority classes or oversampling the minority classes [315]. Virtual sample generation and *Synthetic Minority Over-sampling Technique* (SMOTE) are oversampling techniques. Both can be used to generate synthetic instances for the minority class. Virtual sample generation creates samples by transformations like rotation, scaling, translation, noise addition, or mega trend diffusion [347]. SMOTE interpolates between existing minority class instances and their nearest neighbours [279, 117, 68]. Sharma and Gosain [304] provided a comparison of different SMOTE variants. A common problem encountered with oversampling is that no (real) new information is added to the data set, which can lead to overfitting [315]. Undersampling removes majority class instances, but this usually leads to a substantial loss of information in small data sets [117].

Other generative models exist, such as normalising flows, diffusion networks, or GANs, but they require sufficiently large data [49]. Rauschenberger and Baeza-Yates [284] recommend not to use augmentation methods when the imbalanced data have high variances, as the newly added data are unlikely to adequately represent the class variances.

Feature Augmentation Challenges

If not applied correctly, data and feature augmentation can degrade rather than enhance model performance. This issue often stems from the fact that not all augmentation methods are universally effective. Some methods may introduce noise and artefacts that obscure the underlying data patterns, while others can exacerbate the risk of overfitting [166, 258]. The lack of standardised guidelines for the optimal level of augmentation makes it difficult to determine the appropriate level for a given data set [210]. Additionally, there is a lack of robust metrics for assessing the quality of augmented data, further hindering the evaluation process [258]. Augmentation can also lead to longer training times and higher computational costs, adding another layer of complexity to its implementation [166, 258].

3.4.6 Feature Normalisation

Feature normalisation is a method to adjust the scale and distribution of features so that they are on a common scale or within a specified range. This process ensures that different features contribute equally to ML (not all models require normalised features) [60, 232, 160, 75]. Huang et al. [160] classified normalisation techniques into centering, scaling, decorrelating, standardising, and whitening. For example, a common normalisation technique is min-max scaling to values between 0 and 1.

Normalisation can be challenging with small data sets that lack variability but contain outliers. Outliers are data points that deviate substantially from the norm and can distort the effectiveness of normalisation methods, leading to biased trans-

formations. However, according to Goldstein and Uchida [134], removing outliers can have a negative impact on model performance and does not necessarily guarantee an improvement in ML accuracy, as outliers are usually averaged out during feature extraction. In contrast to normalisation, standardisation (e.g., Z-score) is less sensitive to outliers, as data is centred around the mean and scaled by the standard deviation. However, both methods can be affected by skewed distributions, as some statistical techniques and ML methods assume or perform better with approximately normal data. For example, distance-based algorithms such as k-nearest neighbours are sensitive to feature scales [60, 232].

3.5 Step 5: Machine Learning

Step 5 of the Fatigue Recognition Chain addresses the basic concepts and challenges associated with the application of ML. In particular, fitting an ML model to a small data set which presents unique challenges, including the risk that the model may not generalise. ML is the field of study that gives computers the ability to learn without being explicitly programmed [291]. It draws on concepts from several scientific disciplines, including linear algebra, optimisation problems, probability theory, statistics, and artificial intelligence [179].

3.5.1 ML Method Selection

The choice of which ML method to use is often a trade-off between computational complexity and recognition performance, where the trade-off is influenced by the specific type of activity and the complexity of the feature space being analysed. Other considerations include latency requirements, online processing capabilities, adaptability, and available computing and memory resources [57]. Forman and Cohen [122] showed that feature and model selection are related tasks, and that visualisation of different regions of the learning surface is critical to finding the

optimal combination. Table 2.15 in Section 2.4.2 shows the ML methods commonly used in the related works. Similarly, Kokol et al. [202] found that the most common ML methods used on small data are SVM, DT, RF, CNN, and transfer learning.

Li et al. [231] studied different ML methods in the context of small data sets. They found that SVM often struggles with outliers and noise in the training data. *k*-NN can suffer from reduced precision and classification failure due to an inappropriate value of *k*. ANNs are prone to falling into local minima, have long training times, and the number of hidden layers and nodes is challenging to determine. Statistical learning theory and bootstrap methods may not effectively separate test and training sets, leading to value errors. Bayesian methods can handle incomplete data sets and learn causal relationships between data, while deep neural networks are more suited for perception tasks.

One of the most widely used working assumptions for the design and analysis of ML is the i.i.d. (independent and identically distributed) assumption. The i.i.d. assumption states that the samples in a data set are independent of each other and come from the same probability distribution. Some ML methods are specifically designed for i.i.d. data, such as SVM, *k*-NN, DT, RF, NB, and LR. There are also several ML methods specifically designed to handle non-i.i.d. data, especially time series, such as CNN, LSTM, and RNN. Time series is not i.i.d. because it consists of temporally ordered (consecutive) data points [179] as well as fatigue data due to time-dependent behavioural changes [110].

3.5.2 ML Strategy

A variety of ML strategies exist, such as supervised, unsupervised, semi-supervised, self-supervised, transductive inference, online, reinforcement, active, and transfer learning [128, 283, 254, 237]. In supervised learning, the training set contains target variables (i.e., labels). In unsupervised learning, the training data is unlabelled. Semi-supervised learning combines labelled and unlabelled samples in the training

set, often due to the time and cost associated with labelling the data. A survey of unsupervised and semi-supervised methods with regard to small data can be found in Qi and Luo [283]. Self-supervised learning generates a fully labelled data set from an initially unlabelled one. In reinforcement learning, an agent observes the environment, selects and performs actions, and receives rewards (or penalties) in return, ultimately learning the most rewarding strategy (policy) independently [32, 128]. Transfer learning trains a model on one task as a starting point for a related task, which can be particularly useful when there is limited data available for the new task [237]. Supervised, semi-supervised, and unsupervised learning are commonly employed in related works, as shown in Table 2.15 in Section 2.4.2, with supervised learning being the most commonly used.

ML can be further divided into incremental (online) and non-incremental (batch) learning. In online learning, models are trained incrementally by processing data points or samples one at a time, allowing for continuous updates as new data becomes available. In contrast, batch learning trains the model with the entire data set at once, without supporting incremental updates. In addition, ML can be classified according to how it generalises to new data: instance-based or model-based learning. Instance-based learning generalises by comparing new instances directly with previously observed instances using a similarity measure. In contrast, model-based learning constructs a model from the training data to make predictions, generating classifications, or quantitative outputs based on input features [128].

3.5.3 ML Training

ML models need to be trained before they can be used. Model training is typically performed by dividing the data set into three distinct sets: training, validation, and test. This approach helps to assess model performance and overfitting [149, 128]. The training set is used to train the model by adjusting its parameters to minimise error. In supervised learning, this data set contains correct labels that guide the

model's learning process. In unsupervised learning, the training data lacks labels, leaving the model to discover the underlying patterns on its own. The validation set is used to monitor and optimise the performance of the model during training. It helps to tune the hyperparameters and regulate overfitting. The performance of the model on the validation set also provides insight into how well it might generalise to unseen data. The test set is used to evaluate the final performance of the trained model. It assesses how well the model generalises to unseen data (of the test set) that was not used during training or validation [128].

According to Hastie [149], determining the optimal ratio for each of the three data sets – training, validation, and test – can be challenging. This decision is influenced by factors such as the signal-to-noise ratio in the data and the overall size of the training sample. If the test ratio is too small, it can lead to suboptimal model selection, while if it is too large, it reduces the data available for the other data sets [128]. In the related works, 10% to 30% of the data was usually allocated to the test set (see Section 2.4.2).

3.5.4 ML Prediction

Once the best performing model has been selected, trained and tuned, it is retrained using all available data, including the training, validation, and test sets, to build a final model for deployment in a specific application. This final model can then process any new feature vector to predict a class (classification) or a numerical value (regression) [316]. Other tasks include ranking, clustering, and dimensionality reduction [254]. To further improve the predictive performance of a model, multiple models can be used, known as ensemble or boosting. These models can be fused either at an early stage (i.e., at the feature level) or at a later stage (i.e., at the classifier level). Common fusion methods include summation, majority voting, and Bayesian fusion [57].

3.6 Step 6: Evaluation

Step 6 of the Fatigue Recognition Chain addresses the assessment and optimisation of ML models through various metrics, validation techniques, and evaluation types.

3.6.1 Classification Metrics

Classification is about predicting discrete categories. Common metrics are [57]:

- *True Positives* (TP): correctly predicted positive samples.
- *True Negatives* (TN): correctly predicted negative samples.
- *False Positives* (FP): incorrectly predicted negative samples.
- *False Negatives* (FN): incorrectly predicted positive samples.

These metrics are used to calculate a range of measures, such as [57, 204]:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Recall/Sensitivity/True Positive Rate: $\frac{TP}{TP+FN}$
- Specificity $\frac{TN}{FP+TN}$
- Precision: $\frac{TP}{TP+FP}$
- F_1-score: $\frac{2 \cdot TP}{2 \cdot TP+FP+FN}$

Precision and recall are inversely related; increasing precision often decreases recall and vice versa [204].

Confusion Matrix A confusion matrix is a tool for evaluating the performance of classification models. It visualises the relationship between the predicted and actual class labels for a data set. Each cell in the matrix represents the number of samples that were predicted to belong to one class but actually belong to another [57].

ROC and PR Curve Other common evaluation methods for classification are *receiver operating characteristic* (ROC) and *precision-recall* (PR) curves, which are primarily designed for binary classification [2, 204, 17, 57] (see also Appendix P).

3.6.2 Regression Metrics

In contrast to classification, a regression model predicts continuous numerical values. Different (distance) metrics can be applied for regression [9], the most common are *mean square error* (MSE) $\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, *root mean square error* (RMSE) $\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, *mean absolute error* (MAE) $\frac{1}{n} \cdot \sum_{i=1}^{n}|y_i - \hat{y}_i|$, *mean absolute percentage error* (MAPE) $\frac{100\%}{n} \cdot \sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})$, and *coefficient of determination* (R^2) $1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where $\hat{y}$ is the predicted value, y the actual target variable, and i the index of the feature vector (see also Table 2.16 in Section 2.4.2).

MSE and RMSE use the Euclidean norm (L2 norm) as the distance measure, whereas MAE uses the Manhattan norm (L1 norm). MSE or RMSE are generally the preferred performance measures for regression tasks because they penalise larger errors more than MAE. However, if there are many outliers, MAE may be a better alternative as it is less sensitive to outliers [128]. While MSE, RMSE, and MAE focus on the magnitude of the errors, R^2 measures the proportion of variance in the target variable that is explained by the features [118]. Since R^2 is expressed as a percentage, it can be used to compare different ML models, especially if the target variables have different units. However, Figueiredo Filho et al. [118] criticised R^2 as a statistical measure with little substantive meaning.

3.6.3 Visualisation

(Interactive) visualisation can be helpful in ML evaluation, for example by visualising data characteristics, data distribution, evolving model predictions, or test errors. Through visualisation, researchers can gain a deeper understanding of how models work and identify relevant data, potential biases, and shortcomings [226].

3.6.4 Optimisation

Metrics are the basis for optimising ML models. The optimisation goal may be maximising a single performance metric or multiple metrics simultaneously, depending on the specific application [57]. According to Géron [128], different training, validation, and test data sets can be used to assess generalisation, and once a model shows satisfactory prediction performance – without significant underfitting or overfitting – further hyperparameter tuning can be performed.

3.6.5 Bootstrapping

Bootstrapping is a resampling technique that creates multiple training sets by randomly sampling with replacement from the original data. ML models trained on these bootstrap samples can be evaluated on the out-of-bag data (observations not included in the specific sample) [73, 128]. While bootstrapping provides insight into model variability [128], cross-validation generally provides a more reliable estimate of generalisation performance by systematically testing different subsets of data. Bootstrapping can be useful with small data sets as it allows multiple uses of each data point or feature vector [152, 73], but it does not add new information.

3.6.6 Cross-Validation

Cross-validation is a resampling technique used to evaluate the performance of ML models. It involves partitioning the data into multiple folds, training the model on a subset of the folds, and validating it on the remaining fold. This process is repeated iteratively, with different subsets used for training and validation in each iteration. By averaging the performance metrics across these iterations, cross-validation provides a more reliable estimate of the model's ability to generalise than a single train-test split [57, 128]. A similar evaluation method is *T3-LOSO* and *T4-LMSO*, where the data points or samples are distributed across the folds according to which subject

they belong to. Furthermore, Manna [244] proposed a cross-validation method that uses both training and test measures to reduce the risk of incorrect classification predictions caused by unfortunate data partitioning – the smaller the data set, the greater the likelihood of such errors.

For k-fold cross-validation, *k* determines the number of folds and thus the size of the folds. A potential risk of partitioning small data sets into *k* folds is that some of the created folds may be biased or skewed because they do not contain all classes or are not evenly distributed. For this reason, some metrics, such as Cohen's kappa and F_1 score, may require special treatment. The following metrics address class imbalances in k-fold cross-validation for the F_1 score [296]:

Micro F_1 score evaluates the overall model performance, regardless of class distribution. However, it can be heavily influenced by the majority class, potentially neglecting the performance on minority classes.

$$\frac{2 \cdot \sum_{c=1}^{C} TP_c}{2 \cdot \sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FP_c + \sum_{c=1}^{C} FN_c}$$

where C is the number of classes, TP_c is the number of true positives for class c, FP_c is the number of false positives for class c, and FN_c is the number of false negatives for class c.

Weighted F_1 score weights each class according to its proportion in the data set which can be useful for imbalanced data sets, where some classes have fewer samples and more weight should be given to classes with larger samples.

$$\frac{\sum_{c=1}^{C} (w_c \cdot \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c})}{\sum_{c=1}^{C} w_c}$$

where w_c is $TP_c + FN_c$: the weight, or proportion of samples, in class c.

Macro F_1 score also evaluates overall performance but weights each class equally. It is useful for assessing performance in individual classes, and penalises models that do not perform well in the minority classes.

$$\frac{1}{C}\sum_{c=1}^{C}\frac{2\cdot TP_c}{2\cdot TP_c + FP_c + FN_c}$$

A division by zero can occur when calculating F_1 scores in k-fold cross-validation if a fold is missing samples from all classes. This is more likely with small data sets. To address this issue, data or feature augmentation can be applied. Another measure is stratified k-fold cross-validation, where the data set is divided into *k* folds, but with the additional constraint that each fold must contain approximately the same proportion of samples from each class [10, 58, 274] as illustrated in Figure 3.12.

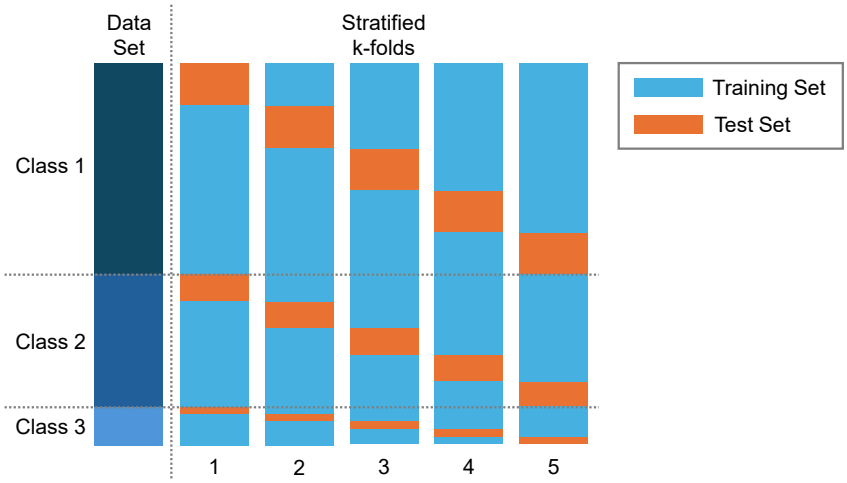


Fig. 3.12.: Stratified k-fold cross-validation of labelled data with three classes.

3.6.7 Evaluation Types

As described in Section 2.4.2, the related works applied one or more evaluation types (*T1-SOLO*, *T2-LNSO*, *T3-LOSO*, *T4-LMSO*) to assess the performance of the trained ML model(s). In *T1-SOLO* and *T2-LNSO* evaluation, k-fold cross-validation is applied on the test set. In *T3-LOSO* evaluation, one subject (i.e., fold) is used as the test set and a model is trained with the remaining subjects, repeating this process until each subject has served as a test set once. The overall performance is usually calculated by averaging the metrics across all cross-tested models [138, 265, 357]. In *T4-LMSO* evaluation, each fold consists of multiple subjects, allowing for different

strategies to generate these folds, such as random subject selection (Monte Carlo) or subject permutation [225]. The advantages and disadvantages of each evaluation type are discussed in Section 6.2.2.

In addition, multiple models can be trained, for example with different subjects, and their results aggregated – typically by majority voting – to make the final decision. This ensemble approach can reduce the dependence on a single test set or subject, potentially improving the robustness and generalisability of the overall model [26].

3.6.8 Generalisation

Generalisation in ML is the ability of a trained ML model to accurately predict results on unseen data which is assessed by testing. Uniform convergence, margin theory, and algorithmic stability are key theoretical tools for understanding generalisation. These concepts help to quantify the relationship between model complexity and the amount of data required to make accurate predictions. While considerable theoretical work has been done, the practical value and applicability of these theories in real world scenarios is still debated [363]. The following section explores factors that influence generalisation in the context of small data, including model complexity, sample variability, and sample complexity, sample size, and sample bias.

Model Complexity

Model complexity influences the ability of a model to generalise to unseen data. The following describes various theoretical and practical aspects for assessing and addressing model complexity.

Hypothesis Space ML methods learn a hypothesis *h(x)*, that processes data *x* and returns a prediction $\hat{y} = h(x)$. Every practical ML method operates within a certain hypothesis space from which the hypothesis *h* is selected (learned). The set of all possible mappings from the sample space to the labelling space, considered

by the learning model, is called the hypothesis space. Figure 3.13 illustrates this concept for a simplified binary classification task with two features. Typically, the hypothesis space of a learning model is an infinite dimensional space. The choice of an appropriate hypothesis space is a critical design decision and is often influenced by domain knowledge [371, 179].

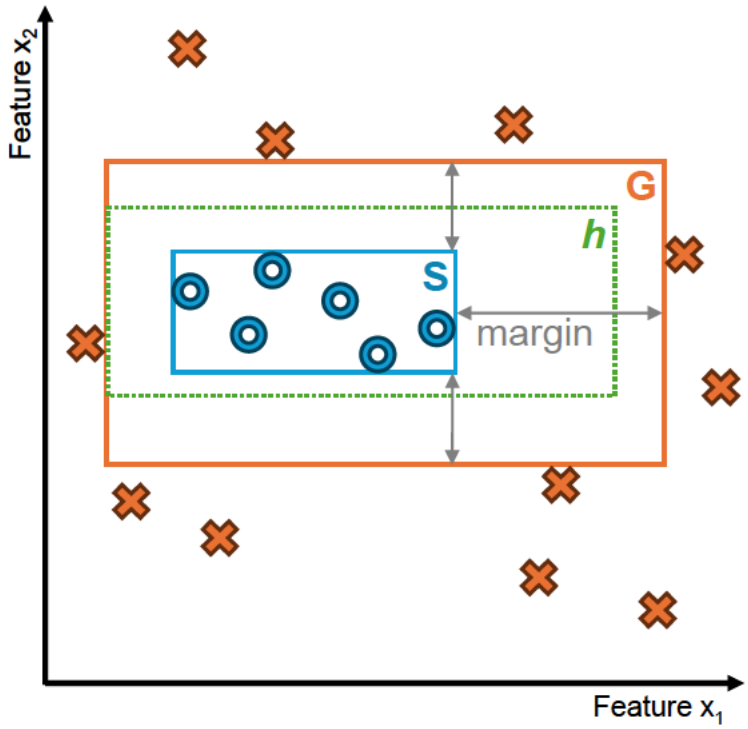


Fig. 3.13.: Box S is the most specific, G the most general, and h is the selected hypothesis. Each blue O and orange X represents a labelled sample.

Quantification of Complexity Theories such Natarajan dimension, Rademacher's complexity, and *Vapnik-Chervonenkis* (VC) dimension provide quantification of the complexity of the hypothesis space [371, 230, 227]. For example, VC quantifies the capacity or complexity of a model by indicating the maximum number of patterns it can memorise or fit without error. The VC dimension is closely related to a model's ability to generalise to unseen data. A higher VC dimension implies a more complex model, that generally requires a larger sample size to learn effectively and avoid overfitting [13, 46].

PAC *Probably Approximately Correct* (PAC) learning theory provides a framework for relating the hypothesis space to sample complexity. The goal of PAC learning is to determine the minimum sample size required to ensure that the learned model has a high probability of being within a specified error margin of the best possible hypothesis. In essence, PAC learning quantifies the number of samples required to develop a model that is probably and approximately correct (within a defined level of accuracy) [13, 46].

Loss Function Due to finite computational resources, an ML method can only consider a subset of all possible hypothesis mappings. For small data, computational resources are less of an issue, but there are concerns about the completeness of the data and whether it adequately covers all relevant possibilities. To learn an appropriate hypothesis from a subset, the quality of a given hypothesis map must be assessed. This is achieved by a loss function, such as the squared error loss $(y-h(x))^2$, which quantifies the difference between the actual data and the predictions made by a hypothesis map. The ML method learns a hypothesis by tuning its internal weights (or parameters) to minimise the average loss, given sufficient samples [179].

Underfitting and Overfitting A fundamental aspect of generalisation is the balance between sample size and model complexity. If the sample size is relatively small, choosing a model with excessive complexity can lead to poor generalisation, commonly referred to as overfitting. Conversely, choosing a model that is too simple may not achieve adequate accuracy, resulting in underfitting [203, 254]. Underfitting occurs when a model is too simple to capture the underlying patterns and relationships in the data, resulting in inadequate model performance. Conversely, overfitting occurs when a model becomes too complex and fits too closely to the training data, including noise and outliers. This results in unstable ML models with low training errors but high testing errors. Ensuring model stability is essential as it generally leads to improved generalisation and consistency [128, 257, 337]. Overfitting with small data is particularly challenging as it involves searching for solutions in a relatively large hypothesis space with probably insufficient data to

provide adequate guidance [204]. Figure 3.14 shows an example of underfitting and overfitting.

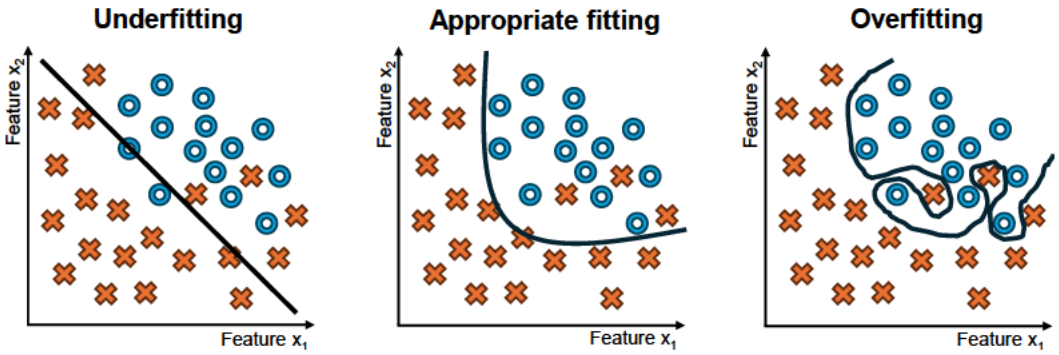


Fig. 3.14.: Examples for underfitting, appropriate fitting, and overfitting.

Possible solutions to overfitting include adjusting the features, collecting more data, reducing noise, and simplifying the model [128]. Simplifying the model is also known as regularisation, which restricts the hypothesis space and penalises complex models with high variance, i.e., when the complexity of a model is very high, regularisation introduces algorithmic tweaks designed to reward models of lower complexity [363]. The regularisation parameter, often denoted as λ, controls the strength of this penalty, balancing the model complexity against its fit to the training data (e.g., the C parameter in SVM). If λ is set too high, only simplistic models are allowed, which may introduce bias [13, 203, 228, 128].

Regularisation techniques are tailored to specific ML methods, for example Lasso (L1) or Ridge (L2) regularisation in linear regression. Their effectiveness often depends on the careful selection of hyperparameters such as learning rate, batch size, or number of epochs [128]. Unlike internal parameters or weights, hyperparameters are not determined by the ML learning algorithm [254]. Instead, they are set by the researcher prior to the learning process. Optimising hyperparameters typically requires tuning strategies, such as cross-validation, to find the right balance between model complexity and performance [13, 46]. In addition, the success of regularisation is highly dependent on data quality – inadequate or unrepresentative

features can reduce the effectiveness of these techniques, leading to less reliable models [361].

Bias-Variance Trade-off The bias-variance trade-off describes the interplay between a model's complexity, the accuracy of its predictions, and its ability to generalise to unseen data, as illustrated in Figure 3.15. Increasing model complexity can reduce

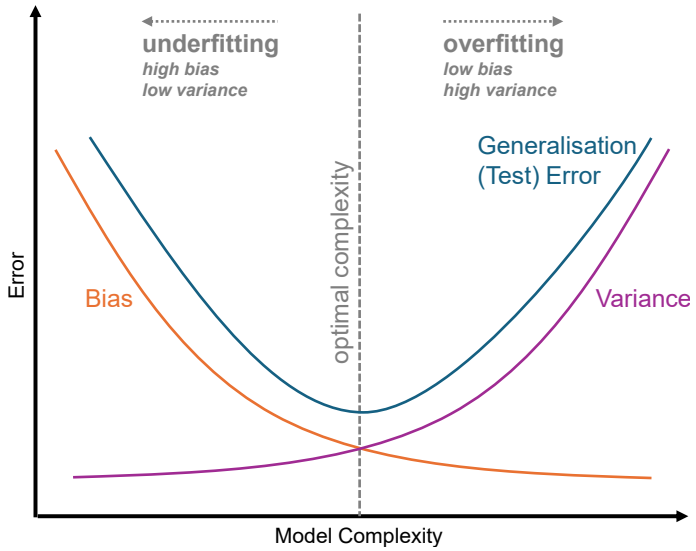


Fig. 3.15.: The classic risk curve of the bias-variance trade-off.

bias but increase variance, while simpler models can have higher bias but lower variance[7]. Bias error is the degree to which the average prediction across all data sets deviates from the desired target. The variance error measures how much the predictions for individual data sets fluctuate around their average, indicating the model's sensitivity to the particular data set. Noise is the unavoidable component of error, independent of the learning algorithm. Generalisation error can be expressed as the sum of bias, variance, and noise: $generalisationerror = bias^2 + variance + noise$ [121].

Radical vs Conservative Generalisation Kong et al. [204] proposed the distinction between radical and conservative generalisation for small data analysis, as shown in Figure 3.16. Radical generalisation creates models that encompass the entire

[7]For very large over-parameterised models, increasing their complexity seems to reduce the test error (double descent phenomenon) [214].

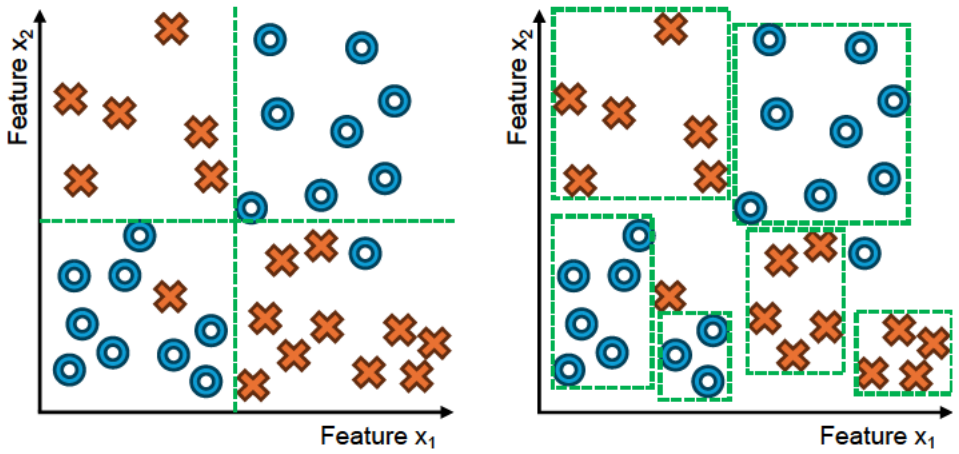


Fig. 3.16.: Left: Radical generalisation (partitioning of data).
Right: Conservative generalisation (selection of data spaces).

data space and groups the data into distinct categories. Conservative generalisation creates models that focus on specific regions with a high concentration of data. This way, broad generalisations are avoided that may not be supported by the data. Conservative generalisation (lattice model) creates a partial model of the data rather than a complete one. This smaller search space reduces the risk of overfitting. However, because the model is partial, it may not cover certain regions of the data space. This can lead to a lower recall rate and the model may miss relevant instances. The challenge is to increase recall and ideally maintain precision [204].

Sample Variability

Sample variability, the spread of data points within a class, has a strong influence on model complexity and thus on generalisation. High variability requires more complex models to capture complex patterns, increasing the risk of overfitting. Conversely, low variability allows for simpler models with reduced overfitting potential [149].

Ensuring the use of a representative data set is critical to the generalisation of ML models. Small sample sizes can be prone to sampling noise, while flawed sampling methods can result in large data sets that are still unrepresentative. Although complex ML models are capable of detecting subtle patterns in data, noisy or inadequately training sets can lead to models identifying patterns in the noise itself

[128, 363]. In practice, data is never homogeneous or has uniform properties such as density, coverage, or instance complexity [25]. It is almost impossible to obtain a data set that comprehensively covers all possible scenarios [62]. Especially small data sets with a low number of subjects usually do not cover all possible variations [67, 56]. A balance has to be found between sample variability, homogeneity, representativeness, and the number of samples. The different aspects that lead to the sample variability are presented in the following.

Intra- and Inter-class Variability The high degree of freedom in human movement leads to considerable variability in how an activity can be performed. This variability is increased by differences in individual body flexibility, body composition and training status [220, 62, 339, 57, 285, 50]. Intra-class variability occurs because the same activity can be performed differently by the same person (in different sessions). Inter-class variability refers to differences in how individuals perform the same activity. Exercises of the same type may be performed by different people with different body movements [57]. The way a person performs an activity may even make it difficult to identify the activity [339]. An inverse challenge is posed by activities (or classes) that are fundamentally different but show very similar characteristics in the sensor data. Such similarities can be resolved by additional cues detected by different sensor modalities [57].

Another challenge is subjectivity, which can vary according to individual perception and familiarity with the scale. This can lead to inaccurate reporting of fatigue levels during exercise, but a subject's familiarity with a scale usually improves with repeated use, leading to more consistent assessments over time [108]. Individual differences in fatigue responses and adaptive capacity also contribute to variability in training load requirements [164]. The training load required for adaptation may vary between individuals [143, 110], but also between sessions for the same individual and is further influenced by various fatigue factors [108] (see Appendix C).

There are several approaches to address intra- and inter-class variability. One approach is to increase the amount of training data, either through more experiments

or through data augmentation. Another approach is to use subject-independent features that are more robust to such variability. For example, features derived from whole-body models rather than low-level signals can improve robustness. However, this requires a trade-off: a highly specific and discriminative feature set may be less effective across individuals, whereas a more generic feature set may offer greater robustness at the cost of lower discriminative power [57]. According to Reid et al. [285], if the intra-class variance of a biometric trait is low, then the trait is said to demonstrate permanence and repeatability. If the inter-class variance is high, then that biometric trait can be successfully used to distinguish between people. Impellizzeri et al. [164] recommended internal load as the primary measure because it better reflects the individual's response to external load. Elshafei et al. [108] applied min-max normalisation to the RPE for the current exercise set to account for subjective differences. Furthermore, a longitudinal approach to data collection may help subjects to become more familiar with the tasks and the scales.

Sensor Variability Variability can also originate from the sensing equipment itself, particularly due to variations in sensor characteristics. This variability can be caused by both internal and external factors. Internal causes include hardware malfunction, complete failure, and sensor drift. External factors can include changes in operating temperature or problems such as loose straps. In addition, some sensors are sensitive to environmental conditions. Wearable devices and sensors can be used differently or positioned in different locations or orientations on the body, which can also contribute to variability [57, 91]. Such sensor issues should be minimised as far as possible, which is usually easier in controlled environments.

Temporal Variability In time series, temporal variability refers to fluctuations and changes that occur over time. This type of variability is particularly evident during exercise due to fatigue and can add complexity to the data. The exercise and the sequence of exercises, including breaks, are other factors that affect variability. For example, if an intensive exercise is performed first, followed by a low-intensity

exercise, there may be after or ripple effects that can manifest as variability in the data [199].

Spatial Variability Spatial variability refers to differences in behaviour or measurements based on the location. Individuals may exhibit different behaviours in different environments. According to Giannakakis et al. [130] and Hussain et al. [162], maintaining constant environmental conditions is challenging due to the wide range of potential factors that may affect individuals, such as noise, temperature, lighting, and air quality.

Sample Complexity

Sample complexity quantifies the relationship between the size of the training data and model performance, with a particular focus on generalisation error. It indicates the amount of data required for a model to perform adequately on unseen data. A model with a high generalisation error typically has a high sample complexity, meaning it requires a larger data set to effectively learn and generalise from the data [13]. Sample complexity is closely related to multiple interrelated concepts: the model complexity, the acceptable margin of error, the probability of failure, and the generalisation error of the model [13].

A practical method for estimating sample complexity is empirical testing, where the size of the data set is gradually increased and the performance of the model is monitored to understand its evolution. This approach provides insight into how the accuracy of the model improves with more data. Alternatively, theoretical methods such as PAC learning bounds can provide guidance [145]. Although these mathematical bounds are often conservative and may not always provide accurate estimates for real world scenarios, they provide a general indication of the sample size required to achieve a desired level of performance. For example, models with high VC dimensions, such as deep neural networks, generally require larger sample

sizes due to their complexity [13, 91]. Similarly, data sets with high variability often require larger sample sizes to effectively capture the underlying patterns [13].

Sample Size

For all learning algorithms trained on sample data, there is a target conflict between the following factors [13]:

1. The complexity of the model that is being adapted to the data.
2. The sample size.
3. The generalisation error for unseen samples.

The relationship between model complexity, sample size, and generalisation is complex. As model complexity increases, performance on training data generally improves, but the risk of overfitting, i.e., poor performance on unseen data, also grows. Conversely, increasing sample size can mitigate overfitting, but only to a certain extent. For example, if the data has been sampled from a straight line and a higher degree polynomial is used, the fitted curve will closely approximate the straight line in regions where there is sufficient training data. However, in regions with sparse data, the polynomial can still deviate substantially from the true pattern, as shown in Figure 3.17. For best generalisation, the complexity of the model should be adjusted to the complexity of the data [13]. However, this requires that there is sufficient representative data for the targeted task available.

Sample Bias

Bias can be broadly defined as a systematic deviation of results or conclusions from the truth, or the processes that lead to such a deviation. It can affect the accuracy, accountability, fairness, and transparency of ML [101]. Sample bias occurs when the sample data used to train an ML model is not representative of the real world population. This can lead to biased or unfair predictions and limit the model's ability to generalise to unseen data. Typical causes include selection and measurement

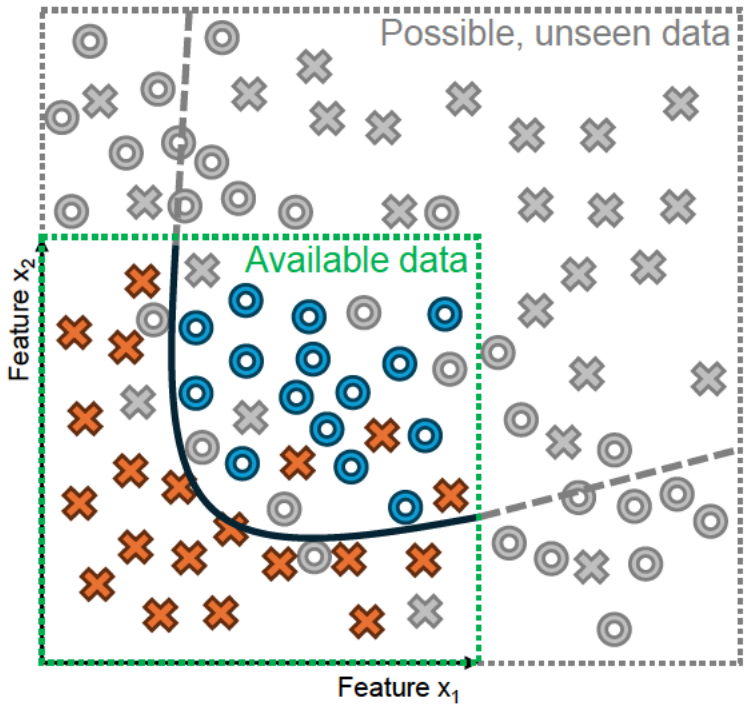


Fig. 3.17.: Fitting an ML model in the absence of data.

bias. Selection bias occurs when the sampling process systematically excludes certain groups or individuals from the data set – or due to inadequate data coverage. Measurement bias occurs when the data collection process is flawed or inconsistent, leading to inaccurate measurements [101, 301, 250].

There are several approaches to mitigate sample bias, such as ensuring that the data is representative of the target population, random sampling to avoid systematic bias, data augmentation to address imbalances or underrepresentation in the data set, and bias detection to identify and correct biases in the data or model [101, 301]. Bias detection in ML requires conscientious efforts at different stages of system development, such as data collection, algorithm design, model training, model evaluation and model deployment. However, a key challenge is to measure bias and fairness in a consistent way, as different stakeholders may have different perspectives, values and preferences on what constitutes bias and fairness and how to quantify them [101]. A comprehensive survey of bias mitigation for ML can be found in Hort et al. [158].

3.7 Step 7: Dissemination

Step 7 of the Fatigue Recognition Chain addresses the documentation and dissemination of the research project, including open data, open source, and open science. Scientific articles can be published as open access, if affordable [123], to increase accessibility – e-Print repositories [75], such as arXiv, should be used to promote open science. In addition, the data and source code should be made publicly available to allow reproducibility, verification, transparency, and collaboration for wider research. Other researchers could use this data to perform secondary analyses and discover new insights by applying different approaches. The data and sources also need to be properly annotated and described, such as column names that reflect the meaning of the data, ideally based on standardised conventions. In addition, the steps in the research process should be documented, including the rationale for each decision made. Version control tools such as Git, SVN or Mercurial can facilitate this process [75, 372]. All of this requires open platforms that should be independent.

3.8 Summary

This chapter introduced the Fatigue Recognition Chain, a structured framework for conducting sensor-based fatigue detection research with ML and small data. The framework consists of seven steps, including foundational characteristics, raw data collection, preprocessing, feature engineering, ML, evaluation, and dissemination. Generalisation was highlighted as part of the evaluation step. Generalisation is the ability of a trained ML model to accurately predict results on unseen data which is assessed by testing. Multiple aspects affect generalisation, such as model complexity, sample variability, sample complexity, sample size, and sample bias. It has been shown that even if a model does not overfit the given data, it may still fail to generalise in a real world application if the training samples lack the relevant

variability or representativeness, which is a key challenge when working with small data.

4 Case Study: Fatigue Detection for Squats with IMU and PE

This chapter presents a case study to demonstrate the application of the Fatigue Recognition Chain framework that has been introduced in the previous Chapter 3. Building on the works of Shi et al. [306], Jiang et al. [176], Wang et al. [343], and Jiang et al. [175], a case study was conducted to illustrate the practical implementation of the framework in the context of fatigue detection for squat exercises with IMU and PE. The results of this case study form an integral part of this thesis by allowing the analysis of ML models trained on small data sets.

4.1 Step 1: Foundational Characteristics

In this step, the foundational characteristics of the case study were established, including the research design and method.

4.1.1 Research Topic & Design

The aim of the case study was to collect RPE, IMU and PE data during squats to train different ML models on small data sets for fatigue detection and to analyse the generalisability of these models. Following the research layers described by Saunders [295], the case study followed a positivist research philosophy and took a deductive approach. An experimental research strategy was adopted to investigate the potential relationship between sensor data and fatigue. This strategy was based on a multi-

method quantitative research design for data collection and analysis. The methods used included multi-sensor data collection, survey scales, data labelling, data mining and hypothesis testing. Data were collected in a multi-session experiment, with one subject per session.

Delimitations: There were certain delimitations to the case study. Firstly, it focused on supervised ML based on subjective RPE as ground truth. Secondly, only squats, a repetitive exercise, were used to induce fatigue in the subjects. Squats was one of the training exercises selected by sports scientists as part of the MoGaSens research project (see below). Thirdly, heart rate monitoring was considered as complement to RPE [45], but preliminary tests showed that heart rate values were highly dependent on the previous activity including the order of the exercises and the rest between them (see Appendix I). In addition, changes in sensor-based heart rate monitoring were delayed in relation to the actual changes in intensity; this delay was not constant and depended on various factors such as age, gender, and fitness level of the individual [274]. For these reasons, heart rate monitoring was excluded from the case study. Fourthly, although lactate testing could provide valuable insights, its cost and the impracticality of frequent blood sampling during short training sessions limited its applicability and it was therefore excluded from the case study.

4.1.2 Research Setting

The research project was conducted at the University of Applied Sciences Hamburg, specifically in the *Creative Space for Technical Innovations* (CSTI)[1] laboratory (see Figure 4.1), where the author was employed as a research assistant during his doctoral studies. The experiments took place on multiple weekdays between 10:00 and 15:00. The experiments were conducted in a laboratory because of the controlled environment with sufficient space (for video recordings and motion tracking) and the availability of the necessary equipment.

[1] `http://csti.haw-hamburg.de/`

Fig. 4.1.: Squat exercises in the laboratory.

4.1.3 Ground Truth

The RPE scale was utilised to collect labels for supervised ML, which is common in related works (see Section 2.4.2), where 46.32% used RPE scales as ground truth. 80% of the related works that used squats as a physical activity also utilised a subjective scale (see Section 2.4.3).

4.1.4 Required Data

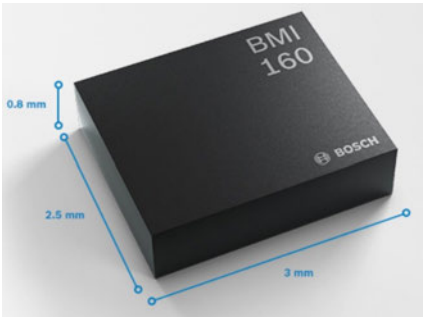
At the time of the case study, there was no labelled and publicly available data on squats. Therefore, new data had to be collected in experiments. As stated by Kluger et al. [199], the technology should be able to examine physiological variables that are hypothesised to be responsible for the fatigue experienced by the target population. The general idea in the case study was to infer fatigue (RPE) from changes in movement patterns similar to studies conducted by Jiang et al. [176], Karvekar et al. [189], Karg et al. [186], and Zhang et al. [367]. Statistical features derived from IMU and PE sensor data were chosen as the main variables. Both sensors capture the subjects' motion data and can also complement each other since

IMU is a wearable sensor and PE is an off-body sensor. As PE has not yet been investigated in this context, a comparison between IMU and PE was also of interest (see Contributions in Section 1.2). Two independent variables would also allow the results to be cross-checked. A brief description and suitability assessment of IMU and PE sensors is provided in the following two sections.

IMU

The need for activity recognition beyond instrumented spaces led to the adoption of body-worn IMU sensors. Since then, there has been considerable research into HAR with unobtrusive, wearable sensors [162].

Definition: An IMU is a wearable sensor for mechanical kinematic sensing (see Appendix J). Modern IMUs are typically based on MEMS[2] technology and com-



Source: https://www.bosch-sensortec.com/products/motion-sensors/imus/bmi160/

Fig. 4.2.: Bosch BMI160 IMU, consisting of accelerometer, gyroscope, and magnetometer.

bine multiple sensors, such as accelerometer, gyroscope, and magnetometer (see Figure 4.2). An accelerometer measures the specific force (in m/s^2), which is the acceleration relative to the inertial frame of reference. This acceleration is the weight experienced by a mass inside the device, as explained by Newton's second law, measured by capacitive, piezoresistive, or thermoelectric sensors. A gyroscope measures the rate of rotation (in $^\circ/s$) caused by changes in attitude relative to inertial space. Gyroscopes use principles such as the Coriolis effect [81]. A magnetometer measures

[2]MEMS stands for Micro-Electro-Mechanical Systems, a technology that combines electrical and mechanical components on a single silicon chip measured in micrometres [178].

the strength and direction of a magnetic field, usually the Earth's [286]. It can act as a compass and help to improve the accuracy and robustness of orientation estimation [47].

Applications: Signals from an IMU can be used to identify physical activity, track the location or motion of objects, track the intensity and frequency of motion, recognise gestures, detect segments, or calculate joint angles. IMUs can also assess the quality of movement [273, 39, 162, 325, 272, 131]. IMUs are often integrated into personal devices, such as smartphones, watches, fitness bands, wristbands, gloves, and rings [39]. According to Tong et al. [325], accelerometers are perhaps the most common wearable sensor in HAR. IMUs have also become a dominant technology for home-based exercise therapy [53].

Benefits: An IMU-based system makes minimal assumptions about the deployment environment and does not require line of sight [233]. An IMU can be ubiquitously deployed in smart devices and often has advantages in power consumption, cost, size, high accuracy, and high sensitivity for HAR [39, 273]. In addition, multiple IMUs can be distributed on the body to improve accuracy [273].

Limitations: IMUs must be worn on the body, which requires robust fixation and precise positioning [39]. Wearing an IMU may not always be practical depending on the task [39, 162]. IMUs also suffer from integrational drift, leading to accuracy problems and accumulated errors over time that must be corrected by constant recalibration [39], which is particularly problematic in precise tracking scenarios. Magnetometers can be ineffective indoors where metallic objects, such as steel frames in walls, can interfere with the sensors [233]. Physical activity can be difficult to distinguish with IMUs [325] as well as detecting precise deviations from ideal movements, such as inappropriate timing of muscle activation. To overcome this, other devices, such as electromyographic or electrocardiographic sensors, must be integrated [272]. In addition, IMUs inherently lack the ability to capture contextual

and social information beyond motion, making the full spectrum of human activity unattainable [325].

Conclusion: The above limitations did not apply to the case study, as only the acceleration and angular velocity signals were required for fatigue detection, with no need for activity recognition or motion tracking.

PE

Historically, HAR has been a focus of computer vision research [39, 162], based on optical signals (see Appendix J), including RGB, depth, infrared, and thermal cameras [62]. Vision-based approaches are unobtrusive [99] and can achieve high recognition accuracy [39]. Vision-based HAR is either marker-based or markerless. Marker-based systems are considered the gold standard in biomechanics and offer the highest accuracy but require wearable markers, specialised equipment, and controlled environments [233]. In contrast, markerless systems eliminate the need for markers and allow for more natural and flexible motion capture [220]. Depth cameras, such as Kinect, leverage technologies like triangulation, time-of-flight, or structured light. They offer a markerless alternative to vision-based HAR approaches. While their affordability and compact size have contributed to their popularity, their accuracy and operating range remain limited in certain environments, particularly in the presence of strong ambient light or reflective surfaces [366].

Definition: Human PE is a markerless, vision-based approach to identify critical body joints or key points within an image or video depicting a person's body. These key points often include joints such as the elbows, wrists, and knees, which are connected to form a coherent structure [209] and can be represented as a skeleton, shape, or mesh model [99].

Methods: Early PE methods relied on predefined models and statistical learning. Since 2013, deep learning techniques, such as CNN, RNN (for temporal information in sequential inputs), graph convolutional networks (ideal for skeleton-based tasks),

and transformers have emerged as the leading techniques [209]. The success of deep learning can be attributed to the abundance of image data, the powerful representational capacity of deep neural networks, and the availability of high-performance hardware [220]. Typical characteristics for PE are top-down or bottom-up, single- or multi-person, and 2D or 3D (mono-view, multi-view, or multimodal) [220]. Top-down approaches first predict the body parts and then compute the poses for each individual (often used for single-person PE). Bottom-up approaches first capture all human body parts and then group these parts to associate them with specific individuals [209]. 2D PE identifies key points within an image plane, while 3D PE extends these spatial coordinates into three-dimensional space [220, 209].

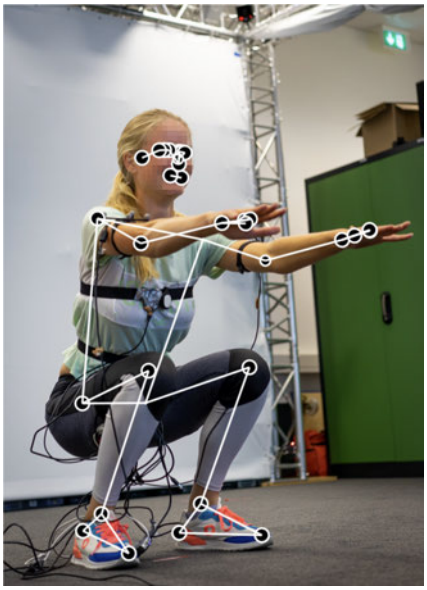


Fig. 4.3.: Skeletal model of a participant performing a squat derived from MediaPipe Pose.

There are many implementations of 2D PE, including MediaPipe Pose, which specialises in exercises. For this reason, MediaPipe Pose was chosen for the case study. It is a lightweight, real-time CNN developed by Google for human PE that extracts 33 key body points for a single person from RGB images. It first identifies human bodies within an image frame, selects the most prominent, and then locates key points on the selected body (see Figure 4.3). MediaPipe Pose estimates the human pose without the need for a graphical model or explicit modelling of the

human body [31, 339]. Appendix N provides more detailed information on how MediaPipe Pose works. Figure 4.4 illustrates a taxonomy for PE adapted from Lan et al. [220] and highlights the characteristics and elements used in the case study.

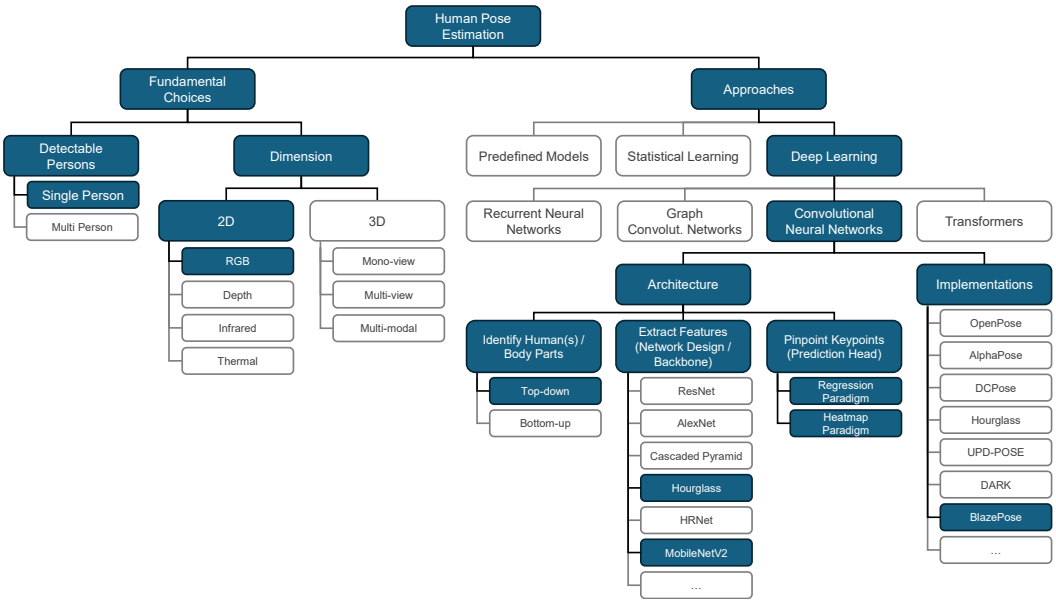


Fig. 4.4.: A taxonomy for human PE, adapted from Lan et al. [220]. The highlighted boxes represent the characteristics and elements utilised in the case study.

Applications: Vision-based approaches cover almost all HAR tasks such as positioning, navigation, body part monitoring, whole body monitoring, individual and group activity recognition [39].

Benefits: PE methods achieve high recognition accuracy [209, 220]. In addition, skeleton tracking can be used to estimate joint angle data [233]. They can also capture contextual information and interaction with people, objects, and the environment.

Limitations: Perspective limitations and occlusion are major challenges for PE. The camera's field of view can partially or completely obscure the subject, while factors such as object interference or self-occlusion can further hinder accurate PE [62, 233]. There is also no semantics to the placement of the camera sensor: a subject can appear in infinite perspectives and scales. The perceived speed of movement is also affected by the distance of the object from the camera [62, 39,

273]. Other challenges include background clutter and moving backgrounds, as well as environmental conditions such as light, shadow, temperature, weather, and air quality [62, 39]. In addition, camera sensors can have shortcomings, such as low video quality, precision, or temperature sensitivity [62, 39, 273]. Real time poses computational challenges due to the high dimensionality of the data and the need for fast processing of high frame rate video streams. The complex nature of human motion, characterised by multiple degrees of freedom per limb, further exacerbates these problems [39, 339]. Privacy is one of the main issues that can lead to discomfort or a sense of intrusion. [62, 39, 273]. Furthermore, a camera cannot follow a moving person without additional equipment.

Conclusion: The controlled laboratory environment minimises computational and environmental limitations, allowing a focused investigation of ML generalisation under optimal conditions. However, this ideal setting may limit the direct applicability to real world scenarios.

Dependence on Body Composition

Acceleration and angular velocity from the IMU can be used for inter-subject comparisons, although these metrics are affected by body composition [289, 7, 350]. This effect can be reduced by homogeneous subject groups. In the case study, the effect of differences in limb length was mitigated by attaching the IMU to the sternum, minimising the influence of individual body proportions. In contrast, PE provides absolute joint coordinates that are highly dependent on camera position and individual body proportions. To allow comparison between subjects, the PE coordinates in the case study were normalised by transforming them to the trajectory-based metrics (i.e., velocity) [366].

4.1.5 Sensor Selection

The following specific sensor models were selected for the case study as they were already available in the CSTI laboratory. A Bosch BMI160 IMU with nine degrees of freedom was employed. It was calibrated once prior to the experiments[3]. In addition, two Logitech c920 webcams were deployed to capture the front and left side of each subject with a fixed sampling rate of 30 Hz at 720p resolution. Marker-based infrared cameras from ART GmbH & Co KG were used for verification purposes only. A comparison between ART and PE can be found in the Appendix K.

4.1.6 Sample Selection

As described in Section 2.4.2, related works recruited an average of 21.1 subjects. A similar cohort was targeted for the case study. However, the number of subjects was gradually increased to examine the effect on classification and generalisation: 20 subjects in the first, 10 in the second, and 18 in the third cohort, for a total of n=48 (32 males and 16 females).

Morris et al. [255] emphasised that variation inevitably affects recognition accuracy and therefore advocated large-scale training. However, when large-scale training is not feasible, it is crucial to minimise variation. Therefore, a balanced set of training data was sought to avoid class imbalances that could negatively affect classification performance [321]. To achieve this, healthy volunteer students from non-sporting disciplines (ecotrophology and computer science) with similar fitness levels (occasional weekly fitness routine) and ages (between 20 and 30 years) were recruited to form a homogeneous group.

[3] https://community.bosch-sensortec.com/t5/Knowledge-base/BMI160-Series-IMU-Design-Guide/ta-p/7376

4.1.7 Ethical Concerns & Consent

Ethical approval for the case study was given by the University of the West of Scotland Ethics Committee. All sessions of the case study were attended by a subject, a sports scientist, and the author. Each session began with the author informing the subject of the general procedure and aims of the study. This was followed by a demonstration of the exercise. A detailed description of the data to be collected and the purpose was given to the subject by the author. The subject then confirmed his or her understanding and agreement by signing a consent form.

4.1.8 Exercise and Sequence

Prior to the start of a session, the subject was equipped with the IMU on the sternum using a chest strap. Meanwhile, the author asked the subject to self-assess their general and current level of fitness and their current RPE. A session consisted of three sets of squats without weights. There was a rest period between each set. Squats were chosen to induce fatigue because, unlike exercises such as push-ups, most healthy people can perform them several times for one minute. Moreover, squats involve the largest muscle group [10]. The subject was then given some instructions for the exercise in terms of foot position (hip width) and arms position (straight forward) – detailed recommendations for performing squats with weights can be found in Comfort et al. [82].

Sensor data was collected throughout the session, including breaks. The subject was asked to report a value between 6 and 20 on the Borg RPE scale every 10 seconds during the exercise. Each exercise lasted 60 seconds, followed by a break of 60 seconds. A session, including the introduction, took approximately 12 minutes per subject. Figure 4.5 illustrates the lab protocol.

In the related works with squats (see Section 2.4.3), RPE were repeatedly collected from a subject, either based on time or after a certain number of repetitions. A 2-minute interval was used in [188, 189], while every 5 repetitions was used in

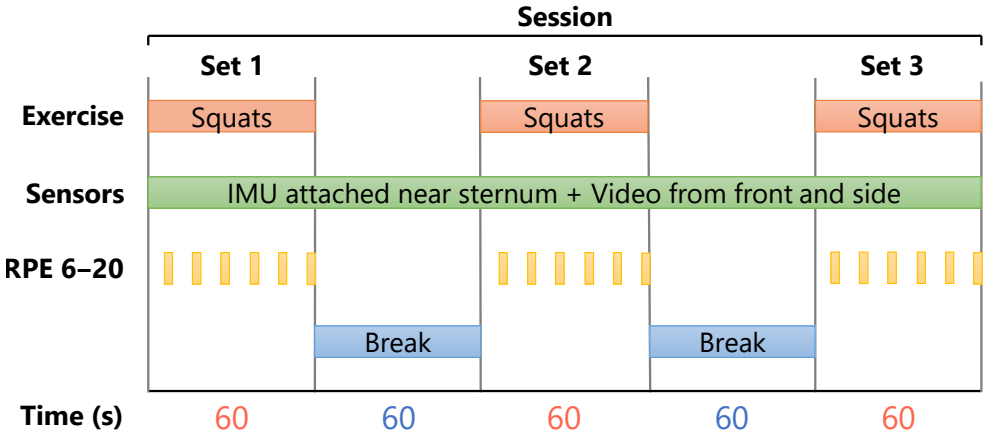


Fig. 4.5.: The laboratory protocol for a session of three sets of squats.

[176, 175, 186], every 12 repetitions in [9, 10], and every 15/30 repetitions in [126]. A time interval of 10 seconds was chosen for the case study. This corresponds to approximately 5–8 squat repetitions. This interval was considered sufficient as initial tests revealed that the same RPE values were repeated more frequently with shorter time intervals.

4.1.9 Unit of Analysis

A squat repetition was chosen as the unit of analysis, specifically the features derived from the time series within the time span of a squat repetition.

4.1.10 Computational Complexity and Storage Requirements

Real time fatigue detection was initially considered. However, given the scope of this thesis and the small data set, offline processing and analysis was adopted. Data collection was performed on a Windows 11 computer equipped with a Core i7-4790K processor, 16 GB of DDR3 RAM, 256 GB SSD, and GeForce GTX 1080 Ti graphics card. Data transformation, feature engineering, ML, and evaluation were performed on a separate Windows 11 computer with Python v3.12 and MATLAB 2024a installed. This machine featured an AMD Ryzen 7 5800X processor, 32 GB of DDR4 RAM, 2 TB SSD, and GeForce RTX 3080 graphics card.

4.2 Step 2: Raw Data Collection

In this step, the raw data from IMU and PE were collected for the case study.

4.2.1 Sampling Rates

Sampling rates of 50–100 Hz were suggested by Trimpop et al. [328] as suitable for the detection of activity and vital parameters. In the related works (see Table B.1 in the Appendix), the sampling rate for IMU ranges from 20 to 1125 Hz, with an average of 189.5 Hz and a median of 100 Hz. In the case study, a sampling rate of 200 Hz was used for the IMU. The webcams had a fixed sampling rate of 30 Hz at 720p resolution – this is lower than recommended, but sufficient for squats as the movements are slow compared to, for example, badminton or boxing.

4.2.2 Data Storage

The data collected for each subject was organised into five *comma-separated values* (CSV) files and one mp4 file[4], each covering a specific aspect of the data collection. The CSV files took up approximately 80 MB of disk space per subject, while an MP4 file took up approximately 200 MB. Table 4.1 gives an example of the file organisation for subject ID 1. Each subject was assigned a unique random ID, which is included in the filename. All filenames followed the same naming convention: *id_datetime_{imu/pe-side/pe-front/sync/borg}.csv*.

Tab. 4.1.: Example of all files stored for the subject with ID 1.

1_2022-11-21_12-47-41.mp4
1_2022-11-21_12-47-41_borg.csv
1_2022-11-21_12-47-41_imu.csv
1_2022-11-21_12-47-41_pe_front.csv
1_2022-11-21_12-47-41_pe_side.csv
1_2022-11-21_12-47-41_sync.csv

[4]Although a time series database would have been beneficial for sharing data between researchers, it was not considered for this PhD project.

Raw IMU data was transferred via a RS232 serial interface using a USB v2.0 cable to minimise data loss. Data was collected with HTerm v0.8.6[5]. Table 4.2 illustrates the information stored in an IMU raw file, including timestamps as well as accelerometer and gyroscope values for the x, y and z axes.

Tab. 4.2.: Example of a CSV file containing raw IMU data.

timestamp_ms	hwTimestamp	accX	accY	accZ	gyroX	gyroY	gyroZ
0	2843910	0.19464111328125	-0.849609375	0.42156982421875	-0.426829268292683	-0.609756097560976	-0.853658536585366
4	2844015	0.19464111328125	-0.849609375	0.42156982421875	-0.426829268292683	-0.609756097560976	-0.853658536585366
8	2844120	0.19659423828125	-0.845458984375	0.42767333984375	-0.853658536585366	-0.609756097560976	-0.853658536585366
...	...	...	...	...	...	...	...

Videos were recorded using Logitech Capture v2.02.155[6] in picture-in-picture mode (front and side view), as shown in Figure 4.6. Since the front and side videos were combined into a single MP4 file by the software, no synchronisation was required. All recorded videos were anonymised with deface v1.1.1[7], a Python-based command line tool for video anonymisation using facial recognition.

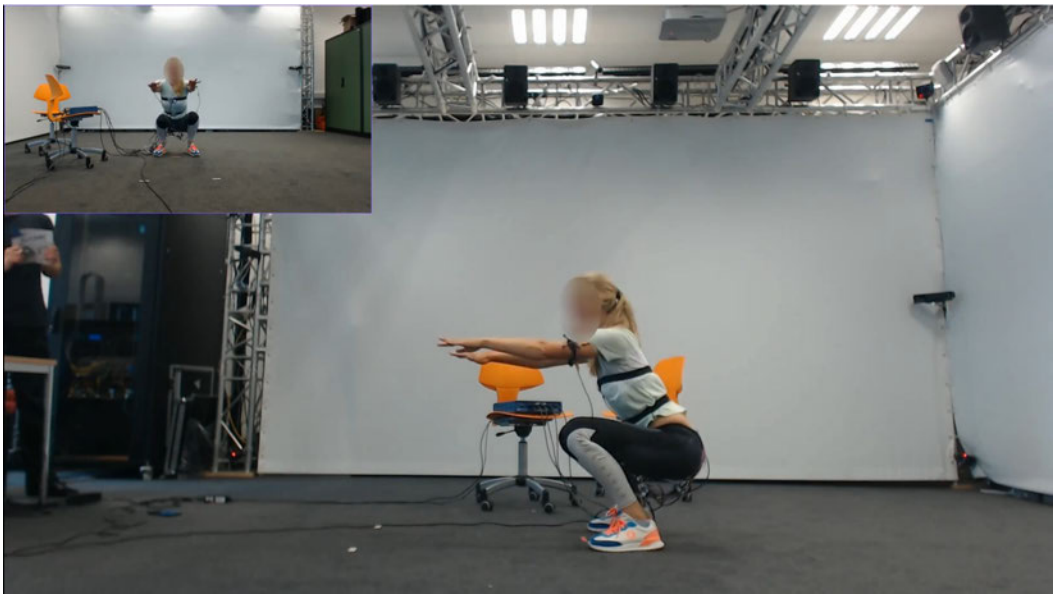


Fig. 4.6.: Picture-in-picture video recording of squat exercises.

[5] https://der-hammer.org/
[6] https://www.logitech.com/en-us/software/capture.html
[7] https://pypi.org/project/deface/

4.2.3 Labelling

RPE from the subject were manually recorded during the exercise in a corresponding borg file. Table 4.3 shows an example where the "label" column contains the only collected data – the other columns contain constant values that have been added to quickly find specific values algorithmically.

Tab. 4.3.: Example of a borg file.

subjectId	setNumber	timeBlock	from	to	label
1	1	1	0	10000	11
1	1	2	10000	20000	12
1	1	3	20000	30000	13
1	1	4	30000	40000	14
1	1	5	40000	50000	14
1	1	6	50000	60000	13
1	2	1	0	10000	11
1	2	2	10000	20000	12
1	2	3	20000	30000	12
1	2	4	30000	40000	13
1	2	5	40000	50000	14
1	2	6	50000	60000	15
1	3	1	0	10000	13
1	3	2	10000	20000	14
1	3	3	20000	30000	14
1	3	4	30000	40000	15
1	3	5	40000	50000	15
1	3	6	50000	60000	15

4.2.4 Synchronisation

For each subject a sensor sync file was created to store offset values in milliseconds to synchronise the different sensors[8]. These offset values referred to the beginning of the first frame of the corresponding video footage for each subject (see Table 4.4).

Tab. 4.4.: Example of a sync file with offset values used for sensor synchronisation.

Exercise	Value_ms
artOffsetInMs	0
hamesOffsetSternumInMs	-725
hamesOffsetBellyInMs	unused
mbientlabsOffsetInMs	unused

[8] Additional sensors for collecting biosignals and another IMU sensor from MBIENTLAB were used in preliminary studies

The offset values for the first 10 subjects were determined manually by comparing the initial acceleration pulses of the PE and IMU signals on a graph: As each subject started in a standing position, the sensor signals were an almost straight line until the subject initiated the exercise. For the remaining 38 subjects, a macro recorder[9] was utilised to initiate recording for all sensors simultaneously, with sensor connections established beforehand to minimise latency. A comparison between the macro and manual synchronisation gave similar results, with an average difference of less than 10 milliseconds, which was considered acceptable given the IMU sampling rate of 200 Hz (i.e., one reading every 4 ms). Figure 4.7 shows the synchronised IMU and PE signals. Albert et al. [10] used a similar approach: they calculated the acceleration in the vertical axis of the Kinect marker and cross-correlated it with the IMU acceleration data.

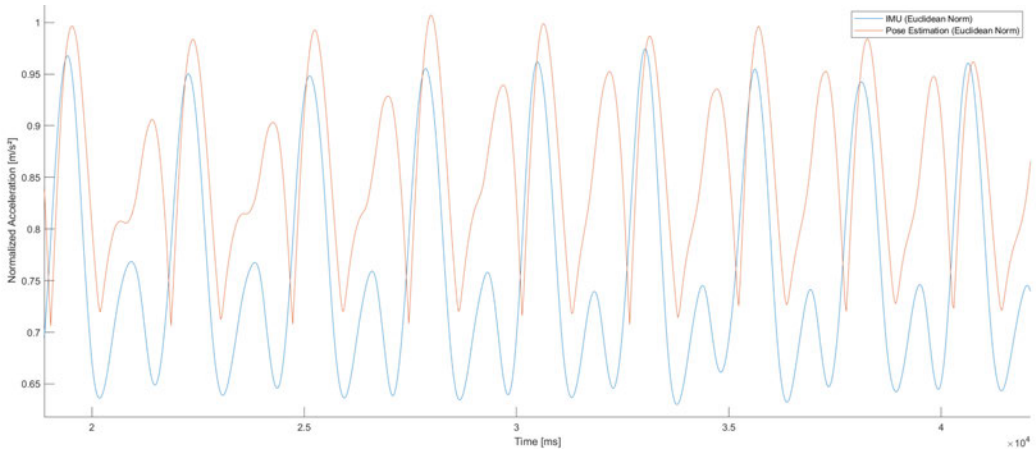


Fig. 4.7.: Synchronised IMU and PE signals through cross correlation.

4.2.5 Data Integrity Verification

Some of the collected data sets were corrupted due to hardware malfunctions caused by computer crashes, disconnections, sensor failures, incorrect camera perspectives, or failures to start recording. Other problems included premature termination of the exercise and incorrect execution by the subject. In the former case, two subjects

[9] `https://www.macrorecorder.com/`

terminated the session prematurely after completing only two sets of squats. In the latter case, one subject's squats deviated considerably from the prescribed squat, resulting in a session reset after corrective instructions were given. A data set (IMU, side PE, or front PE) was considered corrupt if, for any reason, it did not record all the data in a set of exercises. To verify this, each of the recorded IMU and PE signals were plotted on a graph and visually inspected for anomalies such as flat lines, missing data, and unusual signal patterns. As a result, seven sets were considered corrupt and discarded for both IMU and front PE, while three sets were discarded for side PE.

4.3 Step 3: Data Transformation

In this step, the raw IMU and PE data were preprocessed and segmented into individual squat repetitions.

4.3.1 Preprocessing

Different measures were required to preprocess the IMU and PE data. Therefore, both processes are described separately in the following sections.

PE

Additional calculations were required for the PE data to ensure that the measurements were independent of individual body proportions. For this reason, the joint coordinates were converted into joint velocities and joint angles.

Joint Coordinates Extraction The joint coordinates were extracted from the recorded video using MediaPipe Pose v0.9.0.1[10] with default settings. A Python script[11] was

[10] https://developers.google.com/mediapipe/solutions/vision/pose_landmarker

[11] https://colab.research.google.com/github/googlesamples/mediapipe/blob/main/examples/pose_landmarker/python/%5BMediaPipe_Python_Tasks%5D_Pose_Landmarker.ipynb

run separately for the side (PE-Side) and front (PE-Front) videos. Since the videos were recorded in picture-in-picture format, it was necessary to ensure that the PE algorithm selected the correct image. For this reason, the PE-Front videos were cropped from 0 to 383 pixels on the x-axis and 0 to 673 pixels on the y-axis, so that only the front view was visible[12]. To extract the PE-side joints, the top left PE-Front image was blacked out by reusing the same image coordinates as for the cropping. Figure 4.8 visualises the extracted joints.

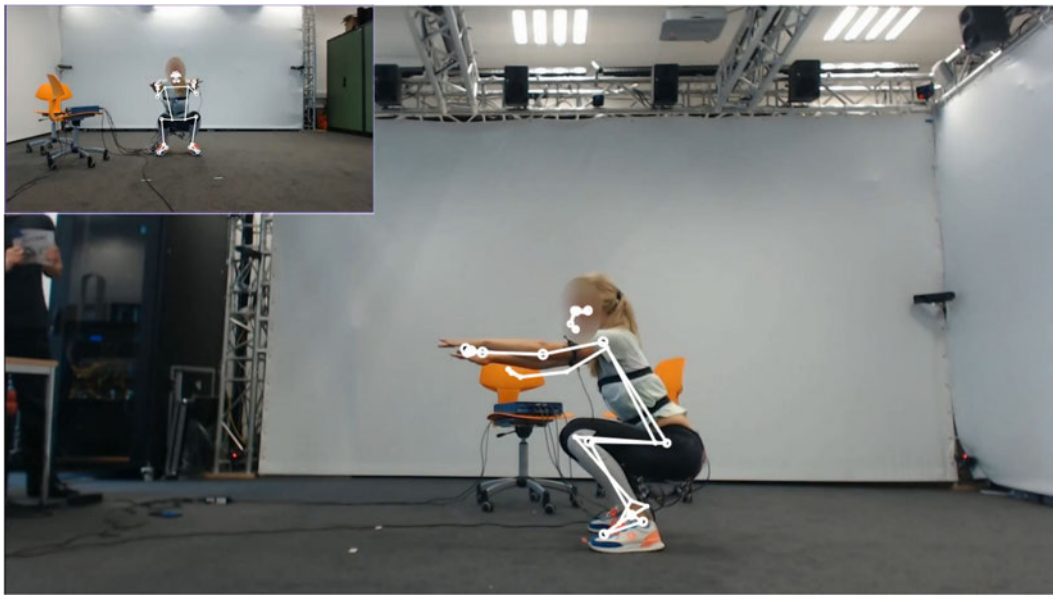


Fig. 4.8.: The joint coordinates for the front and side video were extracted separately.

An extracted data file for PE-Side and PE-Front consisted of 70 columns: the timestamp in milliseconds starting from zero, the absolute x-, y-, and z-coordinates for the nose[13], as well as the absolute x-, y-, and z-coordinates for the left and right shoulder, elbow, wrist, thumb, pinky, index finger, hip, knee, ankle, heel, and foot index. The extracted joint coordinates were normalised by MediaPipe Pose between 0 and 1. MediaPipe Pose could also extrapolate z-coordinates from a 2D image, but this feature was considered less accurate and not used.

[12]The size of the picture-in-picture was set by drag&drop once and used for all recordings.
[13]Due to video anonymisation the extracted nose coordinates were considered unusable.

Joint Velocities Calculation Absolute joint coordinates were not suitable for ML due to differences in body proportions between subjects [9]. One possible solution considered was to normalise the coordinates based on a reference point, such as the central hip joint [5]. However, this approach relies on accurate detection of the reference point. Instead, the velocity, a trajectory-based metric [366, 282], was derived from the joint coordinates. The joint velocities were computed from the extracted joint coordinates over successive frames. Following Aguirre et al. [4], coordinate deltas were computed for the x- and y-axis: $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_i = y_{i+1} - y_i$. The Euclidean distance $d_i = \sqrt{\Delta {x_i}^2 + \Delta {y_i}^2}$ was then computed. Given t_i, the velocity $v_i = \frac{d_i}{t_{i+1} - t_i}$ was obtained in an arbitrary unit per millisecond. The joint velocity was added as an extra column to the respective PE file. This was done for each joint.

Joint Angles Calculation Joint angles were computed based on the positions of two adjacent joints. For example, to compute the left hip angle, the coordinates of the left shoulder a_x, a_y, left hip b_x, b_y, and left knee c_x, c_y were used. The angle β in degrees was determined by the following equation:

$$\beta = \left| (\arctan 2(c_y - b_y, c_x - b_x) - \arctan 2(a_y - b_y, a_x - b_x)) \cdot \frac{180}{\pi} \right|$$

Figure 4.9 illustrates the computed angle of the left hip. The computed angles for all joints were added as extra columns to their respective PE files.

IMU

The IMU data required unit conversion. Additionally, a certain level of cleansing and interpolation was necessary due to missing sensor values.

Unit Conversion The raw IMU data were first converted into acceleration m/s^2 and angular velocity $^\circ/s$ using the following calculations: $acc = rawSensorValue/16.384$ and $gyro = rawSensorValue/16.4$. The divisors represent the sensor sensitivity as specified in the documentation[14] of the IMU.

[14] https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bmi160-ds000.pdf

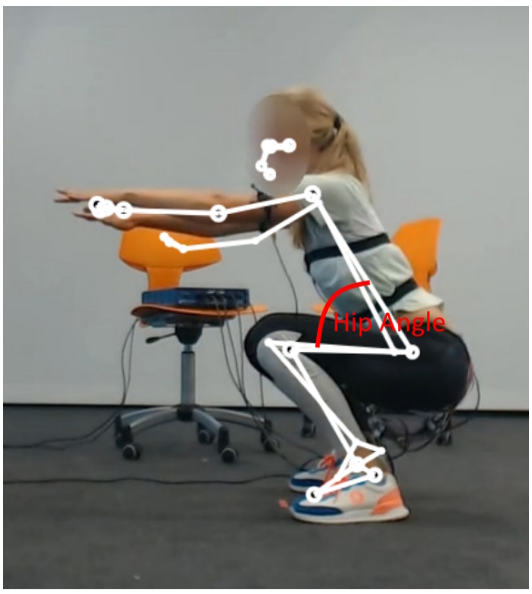


Fig. 4.9.: The hip angle was calculated based on the position of two adjacent joints.

Interpolation Since the IMU data occasionally suffered from minor data loss (less than 1%), the "fillmissing" function in MATLAB[15] with the "pchip" method was used to interpolate missing points. This method performed a shape preserving piecewise cubic spline interpolation. Tests with different interpolation methods showed minimal impact on the ML results, probably due to the high sampling rate and infrequent missing data. Nevertheless, interpolation was used to create a continuous time series, allowing subsequent processing methods such as Butterworth filtering. In addition, an 'interpolated' column was added to the respective IMU file to flag interpolated data points that could be used to assess their impact.

Outliers

The MATLAB function isoutlier[16] with the "percentiles" method and a threshold of 0.1 was used to identify outliers in the IMU and PE data. The total outlier rate was less than 0.3%. As these outliers were removed indirectly during the subsequent computation of statistical features, no explicit removal was applied [134].

[15] https://de.mathworks.com/help/matlab/ref/fillmissing.html
[16] https://de.mathworks.com/help/matlab/ref/isoutlier.html

4.3.2 Motion Segmentation

Various methods for segmentation detection have been proposed in the literature, such as minima and maxima searches [233], also referred to as zero-velocity crossing [53], which is based on physical changes and requires minimal computational resources. However, it is prone to over-segmentation [53] and exhibits limited generalisability across exercise variations and individuals [233]. Additionally, zero-velocity crossing is sensitive to parameters such as sliding window length and can be highly affected by data preprocessing techniques.

As the unit of analysis was a single squat repetition, the preprocessed data were segmented into individual squat repetitions. This step involved identifying the start and end points of each repetition. Figure 4.10 shows a typical repetition cycle. Given the clear peaks on the y-axis of the left hip joint coordinates (PE-Side), these data were used as a reference for segmentation detection with zero-velocity crossing, following the approach described in [4]. The segments detected via the PE-Side data could be used to segment the raw data from the PE-Side, PE-Front, and IMU data because all the data were synchronised.

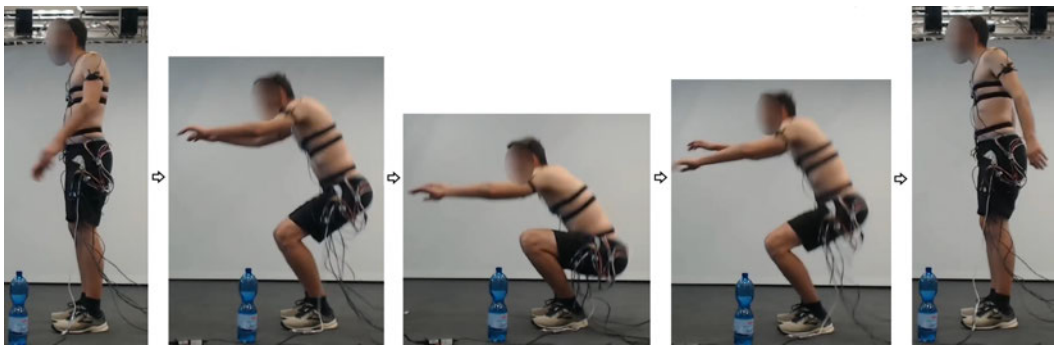


Fig. 4.10.: A complete repetition cycle (i.e., segment) for performing a squat.

Preliminary tests showed that the moving average filter [272] tended to introduce multiple local minima and maxima, complicating peak detection unless extreme filter settings were used. Instead, a third-order low-pass Butterworth[17] filter [38] with a cut-off parameter of 0.06 was applied to the PE-Side joint coordinates (see

[17] `https://de.mathworks.com/help/signal/ref/butter.html`

also Appendix L). These filter parameters were determined empirically to ensure a single local minimum and maximum per repetition for consistent segment detection. As the Butterworth filter is an IIR filter, a constant phase shift correction of -469 ms[18] was applied based on the fixed filter settings and 30 Hz sampling rate of the video recordings (see [76]). The small data set allowed offline segmentation, enabling accurate segment detection using the MATLAB findpeaks[19] function. Table 4.5 lists the empirically determined parameters.

Tab. 4.5.: Parameters used for the MATLAB findpeaks function.

	MinPeakDistance	MinPeakHeight	MinPeakWidth	MaxPeakHeight
Value	800	0.78	10	0.9
Unit	ms	arbitrary unit	arbitrary unit	arbitrary unit

Since the exact number of repetitions was known (from manual counting), a script was implemented to verify that the peak detection algorithm accurately identified this number. If discrepancies were found, missing segments were manually added, and incorrect segments were removed. This approach ensured that all segments were correctly identified. The first and last segments were then removed as they tend

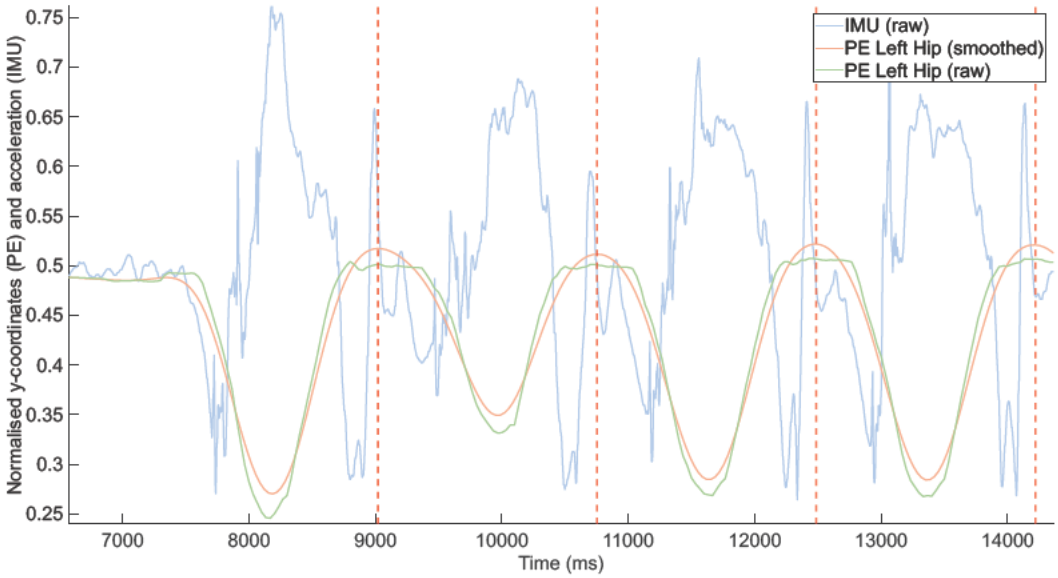


Fig. 4.11.: The vertical red lines show the segmentation slices based on the filtered y-coordinates of the left hip (PE-Side). The first (and last) segment was omitted. Note: The signals have been normalised for visualisation purposes.

[18] For real time applications, other measures would be required to avoid feedback latency.
[19] `https://de.mathworks.com/help/signal/ref/findpeaks.html`

to have irregular motion that distinguishes them from the rest [206]. Figure 4.11 illustrates the segmentation process for the first four repetitions of a set.

All identified start and end timestamps (in milliseconds) of each squat repetition were used to create three new files: imu.mat, peSide.mat, and peFront.mat. These .mat files are in MATLAB's binary format. and can be loaded into memory more quickly than CSV files. Each mat-file stored the respective sensor data from all subjects, with additional columns for subject ID, segment number, set number, RPE (label), and exercise ID (to accommodate future exercises). Table 4.6 provides an example of a segmented mat file for IMU data. Redundant data were included in certain columns to allow flexible data selection during ad hoc experiments, such as plotting data for a specific subject or segment. As RPE were collected every 10 seconds, multiple squat repetitions often shared the same label. Although RPE interpolation was considered [4] (see Appendix H), it was not applied due to potential data distribution alterations.

Tab. 4.6.: Example of an imu.mat file after preprocessing and segmentation with IMU data from all subjects and additional columns for subject ID, set number, segment number, exercise ID, segment duration.
Note: The IMU values have been rounded for visualisation purposes.

timestamp_ms	subjectId	setNumber	segNumber	exerciseId	duration	accX	accY	accZ	gyroX	gyroY	gyroZ	label
9063	1	1	2	1	1833	-0.413	-0.049	0.126	10.548	-54.146	5.914	8
9063	1	1	2	1	1833	-0.401	0.0309	0.146	16.829	-55.365	8.658	8
9063	1	1	2	1	1833	-0.451	-0.111	0.078	20.670	-55.121	11.158	8
9063	1	1	2	1	1833	-0.456	-0.057	0.097	15.914	-48.109	7.926	8
10896	1	1	3	1	1866	-0.475	-0.047	0.153	15.670	-44.695	7.621	8
10896	1	1	3	1	1866	-0.499	-0.057	0.174	13.963	-43.597	7.865	8
10896	1	1	3	1	1866	-0.520	-0.058	0.187	11.951	-43.231	7.926	8

4.3.3 Label Quantity Reduction

Due to small data, the number of RPE labels (originally 14) was reduced to allow more data points to be distributed on each label [212]. Table 4.7 shows how the labels were grouped into different numbers of classes in the case study. These empirically determined groups and thresholds retain an acceptable distribution compared to the original data (see also Section 2.4.2).

Tab. 4.7.: Threshold used in the case study to merge RPE labels.

2 classes	3 classes	4 classes
6–14 / 15–20	6–9 / 10–15 / 16–20	6–9 / 10–11 / 12–15 / 16–20

For each class count, the new labels were normalised to start at 0 and added as alternative labels in extra columns in the segmented mat files, as shown in Table 4.8.

Tab. 4.8.: Example of mapping RPE labels for different numbers of classes in the case study.

timestamp_ms	other columns	label	2 classes	3 classes	4 classes
1020	...	9	0	0	0
14304	...	11	0	1	1
37668	...	15	1	1	2
51176	...	16	1	2	3

4.3.4 Euclidean Norm

In line with related works [343, 140], the Euclidean norm was computed for the accelerometer and the gyroscope data points, respectively, and added to the IMU mat file as new columns.

$$norm = \sqrt{x^2 + y^2 + z^2}$$

The Euclidean norm eliminated the need to identify the axis with the strongest signal and ensured independence from sensor orientation [191]. For example, since the accelerometer detected gravity distributed over all axes, the Euclidean norm effectively aggregated these components into a single value. In addition, using only the Euclidean norm, rather than individual x, y and z axes, reduced the dimensionality of the data. For the latter reason, the Euclidean norm was also applied to the PE data, including the joint velocities and joint angles of the hip, shoulder, and knee. The norm was computed individually for each joint and added as a new column to the respective PE mat file. For PE-Side, the norm was computed as $\sqrt{v_l^2}$ and $\sqrt{\beta_l^2}$, where v_l and β_l represent left-side joint velocities and angles, respectively. For PE-Front, the norm was computed as $\sqrt{v_l^2 + v_r^2}$ and $\sqrt{\beta_l^2 + \beta_r^2}$, taking into account both the left (l) and right (r) joints.

4.3.5 Relevant Data Selection

Some of the collected data was removed as it was not used in subsequent steps. For the IMU data, only the timestamps, labels, and Euclidean norm of the accelerometer and gyroscope were retained for further processing. For the PE data, only the

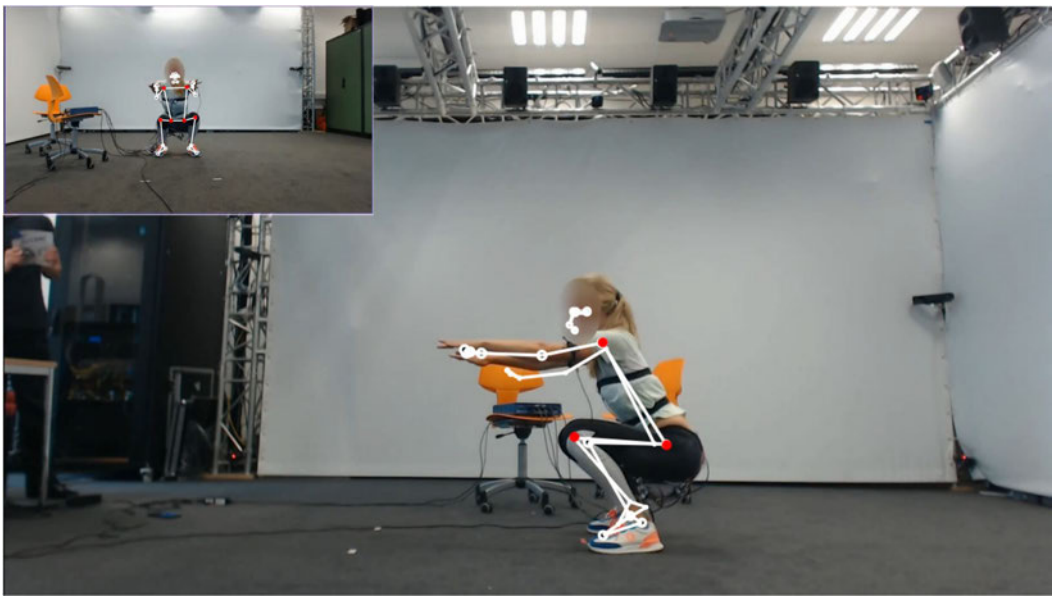


Fig. 4.12.: Only the shoulder, hip, and knee joints (indicated by red dots) were retained for further processing. For PE-Side, only the joints on the left side were kept. For PE-Front, joints on both the left and right sides were kept.

timestamps, labels, and joint velocities and angles of the shoulder, hip, and knee were retained. The PE-Side data set contained only joint data from the left side of the body, while the PE-Front data set contained joint data from both the left and right sides (see Figure 4.12).

4.4 Step 4: Feature Engineering

In this step, relevant features were extracted from the preprocessed, labelled, and segmented data to create a feature set for subsequent ML.

4.4.1 Feature Extraction

Table 4.9 shows the number of non-corrupt segments collected per data source (see Section 4.2.5). Based on these segments, features were computed from each of the raw IMU, PE-Side, and PE-Front data. For the case study, statistical features were adopted based on Guo et al. [140]: median, mean, standard deviation (std), kurtosis, root mean square (rms), skewness, and entropy. These features were computed using corresponding MATLAB functions. The duration of each segment was also added.

Tab. 4.9.: Number of non-corrupt segments collected per data source and the corresponding number of subjects (n).

	IMU	PE-Side	PE-Front
Segments / Samples	3367	3595	3206
n	41	45	41

4.4.2 Feature Normalisation

After calculating the features, the result was a n-dimensional feature vector that was temporarily stored in memory. Each row represented a sample (observation), and each column represented a specific feature. Each feature column was then normalised to a range of 0 to 1 using the minimum and maximum values within the column. This normalisation allowed each feature to contribute equally during ML model training.

4.4.3 Feature Selection

The resulting feature vectors contained 14 feature dimensions for IMU data and 42 dimensions for PE-Side and PE-Front data[20]. Figure 4.10 shows an example of the features computed for the IMU data. The additional columns “subjectId”, “setId”, and “segmentId” were not used as features. Instead, they were used for feature

[20]The data from the left and right joints were combined for PE-Front using the Euclidean norm. As a result, PE-Front had the same number of features as PE-Side.

selection and subject-based model training and evaluation, allowing the construction of subject-specific k-folds. The column “duration” was only used for testing purposes. The column “label” contained the original RPE, while the columns “2 classes”, “3 classes”, and “4 classes” contained grouped labels as described in Section 4.3.3 – one of them was used as a target variable for ML.

Tab. 4.10.: Example of truncated feature vectors for the IMU data. Each row represents a sample (i.e., repetition or segment).
Note: The feature values were rounded for visualisation purposes.

subjectId	setId	segmentId	duration	accEntropy	gyroEntropy	accMedian	gyroMedian	accMean	gyroMean	accStd	gyroStd	$acc_{Kurtosis}$	$gyro_{Kurtosis}$	accRms	gyroRms	accSkewness	gyroSkewness	label	2 classes	...
1	1	2	1833	0.029	0.102	0.025	0.080	0.058	0.068	0.054	0.012	0.024	0.085	0.226	0.171	0.225	0.170	8	0	...
1	1	3	1866	0.034	0.069	0.029	0.061	0.084	0.069	0.077	0.020	0.028	0.069	0.197	0.204	0.197	0.204	8	0	...
1	1	4	1866	0.025	0.065	0.023	0.040	0.029	0.051	0.061	0.011	0.021	0.045	0.221	0.173	0.220	0.172	8	0	...
1	1	5	1799	0.027	0.068	0.023	0.027	0.044	0.024	0.048	0.008	0.021	0.025	0.256	0.150	0.255	0.149	8	0	...
1	1	6	1866	0.028	0.076	0.026	0.024	0.043	0.022	0.056	0.005	0.024	0.022	0.298	0.144	0.297	0.144	8	0	...
1	1	7	1866	0.027	0.068	0.023	0.047	0.030	0.039	0.039	0.011	0.021	0.047	0.270	0.171	0.270	0.171	8	0	...
1	1	8	1866	0.025	0.063	0.022	0.085	0.022	0.086	0.047	0.015	0.020	0.095	0.235	0.187	0.234	0.187	9	0	...
1	1	9	1899	0.034	0.090	0.028	0.071	0.076	0.060	0.048	0.014	0.027	0.074	0.224	0.170	0.224	0.169	9	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Although experiments were carried out with different feature selection methods (see Figure R.1 in the Appendix), including forward and backward selection, these were not pursued further, which is discussed in Chapter 6.

4.4.4 Class Imbalances and Augmentation

Two SMOTE variants were employed to augment the number of segments for the minority classes. When enabled, SMOTE was computed in memory before ML for the selected label (i.e., the selected number of classes). The first technique was the classical SMOTE algorithm without extensions [117]: For each segment, another segment of the same class was identified through a k-NN search (k=3) to generate a new segment. The second technique was a variant of Borderline-SMOTE [165, 279]: For each segment, another segment of the same class and subject was found via k-NN search (k=3). The latter technique produced samples that were closer to the original samples by preserving subject-specific characteristics. In contrast, the classical SMOTE method could create more artificial data samples by combining data from different subjects.

At the time of the case study, there were no specific recommendations in the literature for oversampling percentages in small data sets, consequently an arbitrary

10% of new samples was chosen to create artificial samples for all minority classes. Figures 4.13 and 4.14 show data distributions with and without the application of SMOTE.

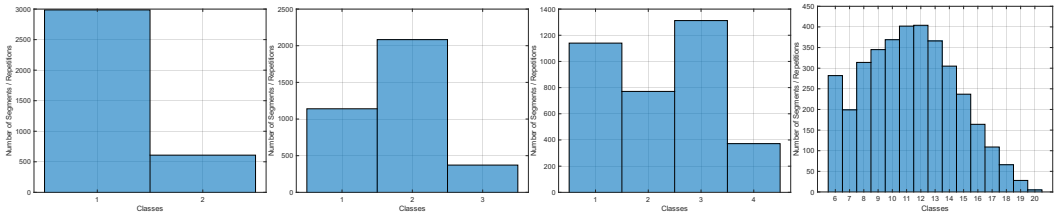


Fig. 4.13.: Distribution of segments for different number of classes without SMOTE.

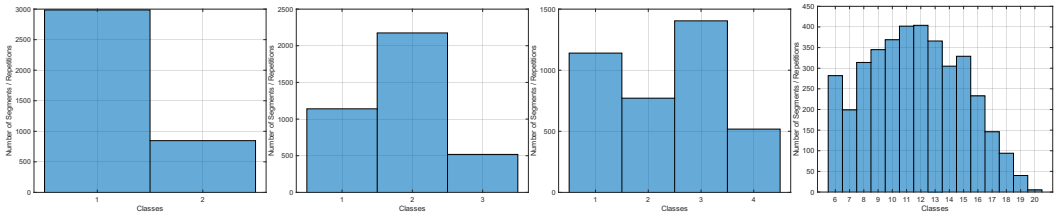


Fig. 4.14.: Distribution of segments for different number of classes with SMOTE.

In the related works, the number of augmented samples was rarely reported (see also Section 2.4.2). Wang et al. [346] collected 483 samples and augmented them to 12558 samples. Albert and Arnrich [9] used a large window overlap of 95% to generate as much training data as possible. Jiang et al. [176] collected samples from 12 subjects and added data from 50 simulated subjects.

4.5 Step 5: Machine Learning

In this step, the feature vectors were partitioned into folds to train different ML models to predict fatigue levels.

4.5.1 Partitioning

MATLAB provides a function cvpartition[21] to partition data into k-folds for cross-validation. However, this function does not take into account the origin of the data, e.g., to leave one subject out. To address this limitation, a custom partitioning function was implemented. This function constructed k-folds so that each fold contained only data from specific subjects, ensuring that there was no overlap between folds. Due to small data, some subject-based folds did not include all classes if a subject's RPE had low variance. These folds were discarded to maintain data integrity. However, a more ideal partitioning function for subject-based k-fold cross-validation should generate stratified folds that include all classes and maintain balanced class distributions across folds. The custom partitioning function was utilised for *T3-LOSO* and *T4-LMSO* evaluation.

4.5.2 Training

Different ML models were trained, with and without oversampled feature vectors, to predict fatigue for each segment. Although various ML methods were used in the related works (see Section 2.4.2), there was no clear consensus on the selection of ML methods. For the case study, the most common ML models were selected, mainly using the default settings of MATLAB. No hyperparameter tuning was performed. Table 4.11 provides an overview of the utilised ML models for classification and their settings. In addition, different regression models with default settings were used. MATLAB's built-in hyperparameter optimisation was disabled due to its high computational cost and limited benefit, according to empirical tests. Given the

Tab. 4.11.: Utilised classification models and their settings.

Support Vector Machine (SVM)	***k*-Nearest Neighbors (*k*-NN)**	**Naive Bayes (NB)**	**Boosted Trees (BT)**	**Artificial Neural Network (ANN)**
kernelFunction: gaussian	DistributionNames: kernel	DistributionNames: kernel	MaxNumSplits: 20	LayerSizes: 100
kernelScale: auto	Distance: euclidean		NumVariablesToSample: all	Activations: relu
	DistanceWeight: equal		NumLearningCycles: 30	Lambda: 0
	k=6		LearnRate: 0.1	IterationLimit: 1000
Standardise: true	Standardise: true	Standardise: true	Standardise: true	Standardise: true

[21] https://de.mathworks.com/help/stats/cvpartition.html

custom partitioning function, the built-in cross-validation function was also disabled. Table 4.12 lists the regression models and their applied settings.

Tab. 4.12.: Utilised regression models and their settings.

Support Vector Machine (SVM)	**Gaussian**	**Tree**	**Ensemble**	**Artificial Neural Network (ANN)**
CrossVal: off	CrossVal: off	CrossVal: off	CrossVal: off	CrossVal: off
OptimiseHyperparameters: none	OptimiseHyperparameters: none	OptimiseHyperparameters: none	OptimiseHyperparameters: none	OptimiseHyperparameters: none
Standardise: true	Standardise: true	Standardise: true	Standardise: true	Standardise: true

4.6 Step 6: Evaluation

In this step, each trained model was evaluated using *T2-LNSO*, *T3-LOSO*, and *T4-LMSO* (see Section 2.4.2). For *T2-LNSO*, a model is trained on 80% of randomly selected segments, with the remaining 20% used as a test set for 5-fold cross-validation. For *T3-LOSO*, a subject-based k-fold cross-validation was applied, where k is equal to the number of subjects used for ML. For *T4-LMSO*, the training set contained approximately 80% of the subjects, with the remaining subjects forming the test set. A 10-fold Monte Carlo cross-validation was applied, with subjects pseudo-randomly selected for the training and test sets, as described in more detail in Section 5.10. After cross-validation, accuracy and ${}_{\text{macro}}\text{F}_1$ scores were computed and averaged to obtain the classification performance. For regression, averaged MAE and RMSE were used to assess the prediction accuracy.

4.7 Step 7: Dissemination

The data could not be published due to lack of informed consent from the subjects. The research results of the case study were published in Jeworutzki et al. [174].

4.8 Summary

This chapter demonstrated the implementation of the Fatigue Detection Chain through a case study and showed how this framework can be used as a guide for conducting similar research. The case study followed a positivist philosophy and adopted a deductive approach using a multi-method quantitative research design. Data collection involved a series of experiments conducted in a controlled laboratory setting at the University of Applied Sciences Hamburg. The case study was based on 48 subjects with similar characteristics.

The experiments for the case study were conducted over several weekdays and included the collection of RPE as a target variable for supervised ML. Squat repetitions are used as the unit of analysis and to induce fatigue, with data collected from IMU and PE to capture changes in movement patterns. The experiments consisted of three sets of squats. RPE were reported every ten seconds during the exercise. Sensor synchronisation was achieved by manually determined offset values and a macro recorder for simultaneous sensor activation. Corrupted data sets with missing data from at least one sensor source were discarded.

Data processing was performed offline. Data transformation included preprocessing steps specific to IMU and PE data. For PE data, faces in the video recordings were automatically anonymised by blurring. In addition, joint coordinates were extracted and converted to relative joint velocity and joint angle values to ensure comparability between segments of subjects with different body proportions. The IMU data underwent unit conversion and interpolation for missing values to maintain a consistent time series. The Euclidean norm was computed to reduce the dimensionality of the features.

Motion segmentation was based on offline peak detection based on hip joint coordinates filtered by a Butterworth filter to identify individual squat repetitions. Segmentation was verified by segment counting. Outliers were removed indirectly by extracting statistical features from each segment. The features were normalised

to a common scale so that each feature contributed equally during ML. The number of labels was reduced to increase the number of data points per label by grouping RPE. SMOTE oversampling was applied to address class imbalances.

Several ML models were trained and evaluated using different settings and cross-validation strategies. A custom partitioning function was implemented to divide the data by subject into folds. Each trained model was evaluated using *T2-LNSO*, *T3-LOSO*, and *T4-LMSO*. For classification, accuracy and ${}_{\text{macro}}\text{F}_1$ scores were computed and averaged over k-folds to assess predictive performance. For regression models, averaged MAE and RMSE were computed. The results have been peer-reviewed and published.

5 Results

This chapter presents an evaluation of the ML models trained in the case study, as described in the previous Chapter 4. First, the general structure of how the results are presented is described. The impact of various factors on model performance is then examined, including classification methods, RPE thresholds, number of classes, evaluation types, data sources, oversampling techniques, feature selection, regression methods, and different number of subjects.

5.1 Result Table Structure

All performance results in this chapter are presented in the form of accuracy, ${}_{\text{macro}}F_1$ score, or confusion matrices (see Section 3.6). Result tables are used to summarise the performance of specific ML models. The results for accuracy and ${}_{\text{macro}}F_1$-score are listed in the upper part of each result table. These values are expressed as normalised percentages ranging from 0 to 1. The “ML Model” column indicates the ML model used for training. The "n" column indicates the number of subjects from which the data was used. The “Data Source” column contains the source of the data signals (IMU, PE-Side, or PE-Front). The “Evaluation Type” column indicates the applied evaluation type. The “Class Count” column indicates the number of classes, i.e., the number of classes into which the RPE20 values were grouped. The “RPE Thresholds” column indicates the thresholds used to group the RPE20 values. The “Feature Count” column indicates how many features were used for the training. This number depends on the source, as the number of signals per source was different. The “Sample Count” column indicates how many samples (observations) were used for training. The “Smote Count” column indicates the number of generated samples.

The “Smote Type” column indicates whether the augmented data was based on each subject separately (intra-subject) or between subjects (inter-subject). The “Test Ratio” column indicates the percentage of test subjects used as test set. The “Test Set Count (k)” column indicates how many test sets (k-folds) were used to evaluate each trained model.

It should be noted that the following results are only comparable to a limited extent as different data sets were used for training and testing, which will be discussed further in the following Chapter 6.

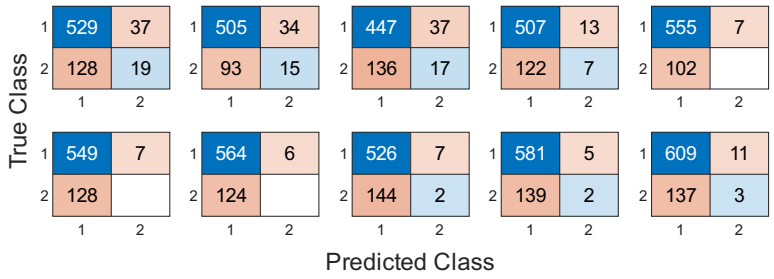
5.2 Classification Models

Table 5.1 compares several ML models, all trained with the same configuration. SVM and DT achieved the highest accuracy with 0.80. *k*-NN achieved the highest $_{\text{macro}}F_1$ score with 0.77. SVM achieved a higher overall accuracy compared to *k*-NN, but lower $_{\text{macro}}F_1$ score. Furthermore, NB achieved slightly higher $_{\text{macro}}F_1$ score than accuracy values.

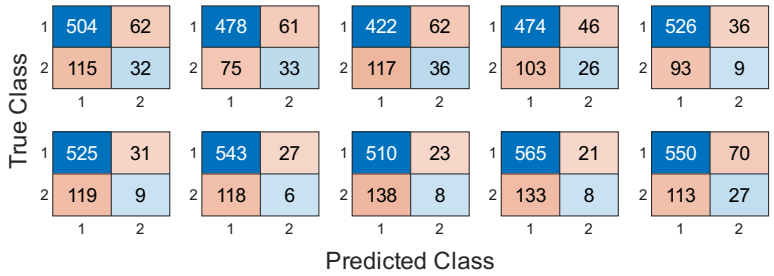
Tab. 5.1.: Results of different ML models trained with the same configuration.

	k-NN Config		SVM Config		ANN Config		DT Config		NB Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.77	0.77	0.80	0.73	0.77	0.74	0.80	0.73	0.72	0.73
Min	0.72	0.70	0.73	0.68	0.70	0.68	0.75	0.66	0.61	0.65
Max	0.81	0.78	0.84	0.77	0.83	0.80	0.83	0.77	0.78	0.80
ML Model	*k*-NN		SVM		ANN		Boosted DT		Naive Bayes	
n	45									
Data Source	PE-Side									
Evaluation Type	*T4-LMSO*									
Class Count	2									
RPE Thresholds	6–14 / 15–20									
Class Distribution	2986 / 609									
Feature Count	42									
Sample Count	3595									
Smote Count	0									
Smote Type	-									
Test Ratio	~20%									
Test Set Count (k)	10									

Figure 5.1 shows the corresponding confusion matrices for SVM and *k*-NN. SVM sometimes did not predict class 2 correctly. In comparison, *k*-NN predicted class 2 more often and more correctly.



(a) SVM Config



(b) *k*-NN Config

Fig. 5.1.: Confusion matrices based on 10-fold cross-validation.
Note: The white tiles without numbers count as zero instances.

5.3 RPE Thresholds

Table 5.2 shows *k*-NN models trained with different RPE thresholds and thus different class distributions. Accuracy and F_1 scores decrease continuously with lower RPE thresholds.

Tab. 5.2.: Results of *k*-NN models with different RPE thresholds and class distributions.

	THRES-15 Config		THRES-14 Config		THRES-13 Config		THRES-12 Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.86	0.83	0.77	0.74	0.64	0.62	0.61	0.60
Min	0.81	0.78	0.72	0.70	0.62	0.59	0.58	0.57
Max	0.88	0.88	0.81	0.78	0.67	0.65	0.64	0.64
RPE Thresholds	6–14 / 15–20		6–13 / 14–20		6–12 / 13–20		6–11 / 12–20	
Class Distribution	3223 / 372		2986 / 609		2681 / 914		2315 / 1280	
ML Model	*k*-NN							
n	45							
Data Source	PE-Side							
Evaluation Type	*T4-LMSO*							
Class Count	2							
Feature Count	42							
Sample Count	3595							
Smote Count	0							
Smote Type	-							
Test Ratio	~20%							
Test Set Count (k)	10							

5.4 Number of Classes

Table 5.3 shows *k*-NN models trained with different number of classes two (C2), three (C3), or four (C4) classes. Thus, each model had different RPE thresholds and class distributions. The accuracy values decreased by approximately 15 percentage

Tab. 5.3.: Results of *k*-NN models with different number of classes.

	C2 Config		C3 Config		C4 Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.86	0.83	0.69	0.52	0.69	0.37
Min	0.81	0.78	0.65	0.48	0.68	0.33
Max	0.88	0.88	0.71	0.59	0.72	0.42
Class Count	2		3		4	
RPE Thresholds	6–15 / 16–20		6–9 / 10–15 / 16–20		6–9 / 10–12 / 13–15 / 16–20	
Class Distribution	3223 / 372		1140 / 2083 / 372		1140 / 771 / 1312 / 372	
ML Model	*k*-NN					
n	45					
Data Source	PE-Side					
Evaluation Type	*T4-LMSO*					
Feature Count	42					
Sample Count	3595					
Smote Count	0					
Smote Type	-					
Test Ratio	∼20%					
Test Set Count (k)	10					

points from C2 to C3 and C4. The accuracy values remained almost the same for C3 and C4. The $_{\text{macro}}F_1$ scores decreased by about 30 percentage points from C1 to C2 and by about 30 percentage points from C2 to C3.

5.5 Evaluation Types

Table 5.4 shows *k*-NN models trained with *T2-LNSO* (T2), *T3-LOSO* (T3), or *T4-LMSO* (T4) evaluation (see also Section 2.4.2). For T2, the test set contained 20% (719) of all samples. For T3, each test set contained the data of one subject and the results were averaged from n test sets (folds). For T4, each test set contained data from 9 subjects (20% test ratio) and the results were averaged from 10 test sets (folds). For T2, the accuracy and $_{\text{macro}}F_1$ score values for T2 were almost the same (0.87–0.88). The range between the minimum and maximum values was 0. For T3, the accuracy and $_{\text{macro}}F_1$ score values were lower than for T2. The range between the minimum and maximum values was: 0.29–0.97 for accuracy and 0.21–0.96 for

Tab. 5.4.: Results of *k*-NN models with different evaluation types.

	T2 Config		T3 Config		T4 Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.88	0.87	0.73	0.67	0.77	0.77
Min	0.88	0.87	0.29	0.21	0.72	0.70
Max	0.88	0.87	0.97	0.96	0.81	0.78
Evaluation Type	*T2-LNSO*		*T3-LOSO*		*T4-LMSO*	
ML Model	*k*-NN					
n	45					
Data Source	PE-Side					
Class Count	2					
RPE Thresholds	6–14 / 15–20					
Class Distribution	2986 / 609					
Feature Count	42					
Sample Count	3595					
Smote Count	0					
Smote Type	-					
Test Ratio	~20%					
Test Set Count (k)	10					

$_{\text{macro}}F_1$ score. For T4, the accuracy and $_{\text{macro}}F_1$ score values were lower than for T2, but higher than for T4. In addition, the range between minimum and maximum was 0.09 for accuracy and 0.08 for $_{\text{macro}}F_1$ score.

5.6 Data Sources

Table 5.5 shows *k*-NN models trained with data from IMU, PE-Side, and PE-Front. IMU achieved similar accuracy values to PE-Front (0.72), but lower average $_{\text{macro}}F_1$ scores (0.67 vs 070). PE-Side achieved the highest average accuracy (0.77) and $_{\text{macro}}F_1$ score (0.74).

Tab. 5.5.: Results of *k*-NN models with different data sources.

	IMU Config		PE-Side Config		PE-Front Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.72	0.67	0.77	0.74	0.72	0.70
Min	0.69	0.60	0.72	0.70	0.62	0.66
Max	0.78	0.77	0.81	0.78	0.78	0.75
Data Source	IMU		PE-Side		PE-Front	
n	41		45		41	
Feature Count	14		42		42	
Sample Count	3367		3595		3206	
Class Distribution	2771 / 596		2876 / 719		2617 / 589	
ML Model	*k*-NN					
Evaluation Type	*T4-LMSO*					
Class Count	2					
RPE Thresholds	3–14 / 15–20					
Smote Count	0					
Test Ratio	~20%					
Test Set Count (k)	10					

5.7 Oversampling

Table 5.6 shows *k*-NN models trained with oversampled data and different oversampling techniques (see also Section 4.4.4). All trained models with oversampled data achieved lower average accuracy and $_{\text{macro}}F_1$ scores. SMOTE-INTRA, which augmented data for each subject separately, achieved the highest average accuracy and $_{\text{macro}}F_1$ score among the oversampled configurations.

Tab. 5.6.: Results of *k*-NN models with different oversampling settings.

	SMOTE-NONE Config		SMOTE-INTER Config		SMOTE-INTRA Config		SMOTE-BOTH Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.77	0.74	0.70	0.64	0.71	0.68	0.62	0.58
Min	0.75	0.70	0.64	0.59	0.66	0.64	0.54	0.51
Max	0.81	0.78	0.74	0.70	0.74	0.74	0.68	0.67
Smote Count	0		360		360		360	
Smote Type	-		Inter-Subject		Intra-Subject		Inter-/Intra-Subject	
Class Distribution	2986 / 609		3285 / 670		3285 / 670		3285 / 670	
Sample Count	3595		3955		3955		3955	
Smote *k*-NN	3							
ML Model	*k*-NN							
n	45							
Data Source	PE-Side							
Evaluation Type	*T4-LMSO*							
Feature Count	42							
Class Count	2							
RPE Thresholds	3–14 / 15–20							
Test Ratio	~20%							
Test Set Count (k)	10							

5.8 Feature Sets

The following classification results were based on different feature sets and different data sources (IMU, PE-Side, or PE-Front).

5.8.1 IMU Features

Table 5.7 shows *k*-NN models trained with different numbers of features based on IMU data. KINEMATIC was based on features from accelerometer data, ANGLE was based on the gyroscope data. KINE-ANGLE was a combination of KINEMATIC and ANGLE. KINEMATIC achieved the lowest average accuracy and $_{\text{macro}}F_1$ scores. ANGLE achieved a slightly higher average accuracy (0.73) than KINE-ANGLE (0.72),

Tab. 5.7.: Results of *k*-NN models with different IMU feature sets.

	KINEMATIC Config		ANGLE Config		KINE-ANGLE Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.64	0.61	0.73	0.65	0.72	0.67
Min	0.57	0.55	0.68	0.57	0.69	0.60
Max	0.75	0.71	0.82	0.77	0.78	0.77
Feature Count	7		7		14	
ML Model	*k*-NN					
n	41					
Data Source	IMU					
Evaluation Type	*T4-LMSO*					
Class Count	2					
RPE Thresholds	3–14 / 15–20					
Class Distribution	2771 / 596					
Sample Count	3367					
Smote Count	0					
Test Ratio	~20%					
Test Set Count (k)	10					

but a wider minimum and maximum range. KINE-ANGLE achieved the highest $_{\text{macro}}F_1$ scores (0.67).

5.8.2 PE-Side Features

Table 5.8 shows *k*-NN models trained with different numbers of features based on data from PE-Side. SHOULDER, HIP and KNEE were each based on the joint velocity and joint angle velocity from the respective joint. KINEMATIC was based on velocity data for shoulder, hip, and knee joint combined. ANGLE was based on joint angle data for shoulder, hip, and knee joint combined. KINE-ANGLE was a combination of KINEMATIC and ANGLE. Of the three joints, SHOULDER achieved the highest

Tab. 5.8.: Results of *k*-NN models with different PE-Side feature sets.

	SHOULDER Config		HIP Config		KNEE Config		KINEMATIC Config		ANGLE Config		KINE-ANGLE Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.78	0.74	0.75	0.70	0.75	0.70	0.78	0.74	0.81	0.78	0.77	0.74
Min	0.72	0.70	0.70	0.63	0.67	0.66	0.75	0.69	0.76	0.73	0.72	0.70
Max	0.83	0.79	0.79	0.75	0.80	0.76	0.81	0.78	0.85	0.83	0.81	0.78
Feature Count	14		14		14		21		21		42	
ML Model	*k*-NN											
n	45											
Data Source	PE-Side											
Evaluation Type	*T4-LMSO*											
Class Count	2											
RPE Thresholds	3–14 / 15–20											
Class Distribution	2876 / 719											
Sample Count	3595											
Smote Count	0											
Test Ratio	~20%											
Test Set Count (k)	10											

average accuracy (0.78) and $_{\text{macro}}F_1$ score (0.74). HIP and KNEE achieved similar values. ML models trained with features from KINEMATIC or ANGLE only, performed

worse than models trained with a combination of both. ANGLE achieved the highest overall average accuracy (0.81) and $_{\text{macro}}F_1$ score (0.78). KINE-ANGLE performed slightly worse than SHOULDER.

5.8.3 PE-Front Features

Similar to PE-Side, ML models were trained for PE-Front. Table 5.9 shows *k*-NN models trained with different numbers of features based on data from PE-Front. Of the three joint, SHOULDER and HIP achieved the highest average accuracy and

Tab. 5.9.: Results of *k*-NN models with different PE-Front feature sets.

	SHOULDER Config		HIP Config		KNEE Config		KINEMATIC Config		ANGLE Config		KINE-ANGLE Config	
	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$	Accuracy	$_{\text{macro}}F_1$
Average	0.77	0.71	0.77	0.71	0.74	0.72	0.75	0.71	0.75	0.70	0.72	0.70
Min	0.69	0.61	0.72	0.64	0.64	0.67	0.70	0.64	0.71	0.63	0.62	0.66
Max	0.82	0.79	0.80	0.76	0.81	0.76	0.80	0.76	0.79	0.74	0.78	0.75
Feature Count	14		14		14		21		21		42	
ML Model	*k*-NN											
n	41											
Data Source	PE-Front											
Evaluation Type	*T4-LMSO*											
Class Count	2											
RPE Thresholds	3–14 / 15–20											
Class Distribution	2617 / 589											
Sample Count	3206											
Smote Count	0											
Test Ratio	~20%											
Test Set Count (k)	10											

$_{\text{macro}}F_1$ scores with slightly different minimum and maximum values. KINEMATIC and ANGLE performed almost the same and achieved slightly lower average accuracy and $_{\text{macro}}F_1$ scores than SHOULDER or HIP. KINE-ANGLE achieved the lowest overall average accuracy and $_{\text{macro}}F_1$-score.

5.8.4 Comparison of IMU, PE-Side, and PE-Front

On average, SHOULDER, HIP and KNEE achieved similar average accuracy (0.76) and $_{\text{macro}}F_1$ scores (0.71). ANGLE results for average accuracy (0.81) and $_{\text{macro}}F_1$ score (0.78) of PE-Side were notably higher than PE-Front (0.75 and 0.70). Contrary to IMU, PE did not achieve the best performance with a combination of KINEMATIC and ANGLE.

5.9 Regression Models

Table 5.10 shows different regression models trained with the same configuration. GAUSSIAN-R achieved the best average MAE of 2.57, while ENSEMBLE-R achieved the best average RMSE of 0.57.

Tab. 5.10.: Results of different regression models with the same configuration.

	SVM-R Config		GAUSSIAN-R Config		TREE-R Config		ENSEMBLE-R Config		ANN-R Config	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Average	2.89	2.78	1.08	2.57	0.88	3.27	0.57	3.32	2.50	2.83
Min	2.79	2.54	0.96	2.38	0.81	2.90	0.55	2.99	2.35	2.55
Max	2.95	3.09	1.14	2.73	0.95	3.60	0.59	3.86	2.79	3.30
ML Model	SVM		Gaussian Process		Binary Tree		DT with LSBoost		Neural Network	
n	45									
Data Source	PE-Side									
Evaluation Type	*T4-LMSO*									
Feature Count	42									
Class Count	14									
RPE Thresholds	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20									
Class Distribution	282, 199, 303, 332, 356, 390, 390, 363, 305, 237, 164, 109, 66, 28, 5									
Sample Count	3595									
Smote Count	0									
Test Ratio	~20%									
Test Set Count (k)	10									

Figure 5.2 exemplarily illustrates the predicted and actual regression values of the GAUSSIAN-R model for the test set that achieved the best average MAE.

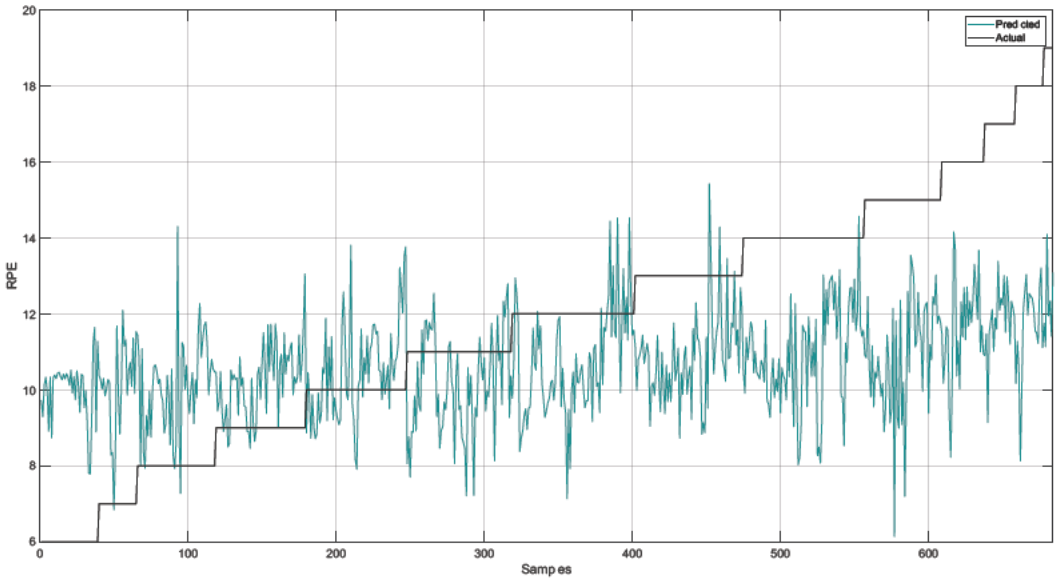


Fig. 5.2.: Gaussian regression with predicted and actual values (RPE per sample).

5.10 Incremental Number of Subjects

This section examines the evaluation results of ML models for an increasing number of subjects. Box-and-whisker diagrams were utilised for visualisation. First, the construction of the test sets and the diagram are briefly explained.

5.10.1 Test Sets Construction

Table 5.11 shows an example of how the first 4 test sets were constructed with $k = 10$ folds for different numbers of subjects (n) to achieve a constant test ratio of about 20%. The aim was to allow more comparable evaluation of the ML models for different n by using training and test sets with data from mostly the same subjects. Experiments also indicated that more folds than $k > 10$ had little effect on classification performance.

Tab. 5.11.: Example of constructing 10 test sets for different n. Each row contains the test sets for a particular n. Each column represents 1 of the 10 test sets (folds). The values in the cells are the subject IDs from which the test data are taken.

	Test Set 1	Test Set 2	Test Set 3	Test Set 4	Test Set 5	Test Set 6	Test Set 7	Test Set 8	Test Set 9	Test Set 10
n=5	39	31	21	27	11	-	-	-	-	-
n=8	39, 31	31, 21	21, 27	27, 11	11, 13	13, 28	28, 34	-	-	-
n=13	39, 31, 21	31, 21, 27	21, 27, 11	27, 11, 13	11, 13, 28	13, 28, 34	28, 34, 24	34, 24, 2	24, 2, 46	2, 46, 40
n=18	39, 31, 21, 27	31, 21, 27, 11	21, 27, 11, 13	27, 11, 13, 28	11, 13, 28, 34	13, 28, 34, 24	28, 34, 24, 2	34, 24, 2, 46	24, 2, 46, 40	2, 46, 40, 18

5.10.2 Box-and-Whisker Diagram

The box-and-whisker diagram[1] shows the accuracy or ${}_{\text{macro}}F_1$ scores on the y-axis and the number of subjects used to train the ML model on the x-axis. Each of these models was tested with 10 test sets (k-folds). The minimum, maximum, mean, and median for accuracy and ${}_{\text{macro}}F_1$ score were computed based on the results of these test sets. The metrics are presented as normalised percentages ranging from 0 to 1 in the diagram. Additional summary statistics across all n results are displayed in the bottom right corner of the diagram. Below this are the constant settings used to train each model.

The blue boxes in the diagram represent the *interquartile range* (IQR), which is the range between the first quartile (Q1, the 25th percentile) and the third quartile

[1] `https://de.mathworks.com/help/stats/boxplot.html`

(Q3, the 75th percentile). This range contains the middle 50% of the data. The red line within a blue box represents the median (the 50th percentile) of the data set. The plotted whisker extends to the most extreme accuracy or ${}_{\text{macro}}\text{F}_1$ score.

For *T4-LMSO* evaluation, the vertical grey dotted lines indicate changes in the number of subjects used for each test set (fold) as described in Section 5.10.1.

The variance coefficients and entropy (min-max normalised) were plotted in the diagram as dashed curves to assess the variability of the data. These metrics were divided into "Kinematic", referring to acceleration for IMU or joint velocity for PE, and "Angle", referring to angular velocity for IMU or joint angles for PE. They were computed on the entire raw data set based on the current number of subjects.

5.10.3 n Models with *T4-LMSO* Evaluation

Figure 5.3 shows the box-and-whiskers diagram with ${}_{\text{macro}}\text{F}_1$ scores for *k*-NN models, all trained with the same settings but with different number of subjects (see Figure S.1 for the accuracy results in the Appendix). The range of results (height of boxes and whiskers) decreased as n increased. The range at n=45 is considerably lower than the range of models with n<38. Furthermore, the range decreased abruptly when the number of subjects in the test set was increased by one. The medians (red lines) converged slowly around 0.72 as n increased. The variance coefficients and entropies increased only slightly with increasing n, except for the angle entropy which showed erratic values.

5.10.4 n Models with *T3-LOSO* Evaluation

Figure 5.4 illustrates the ${}_{\text{macro}}\text{F}_1$ scores using identical parameters, but with *T3-LOSO* evaluation. Unlike *T4-LMSO*, where the range between the minimum and maximum scores decreased as n increased, the range remained relatively constant for the 10 test sets in *T3-LOSO* evaluation. From n=11, however, there was a noticeable increase in the range, which remained almost unchanged up to n=45.

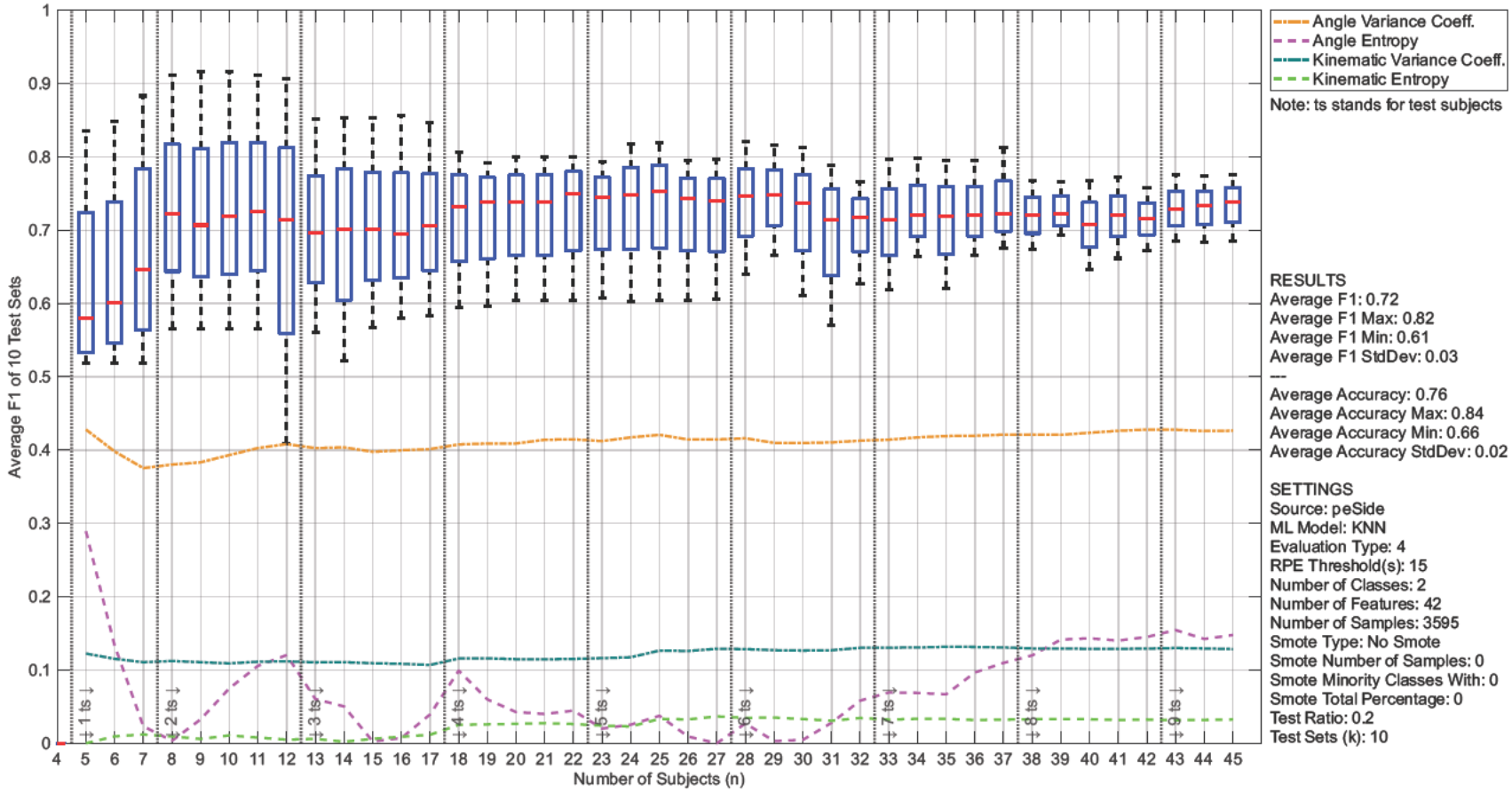


Fig. 5.3.: $_{\text{macro}}F_1$ scores for an incremental number of subjects with *T4-LMSO* evaluation.

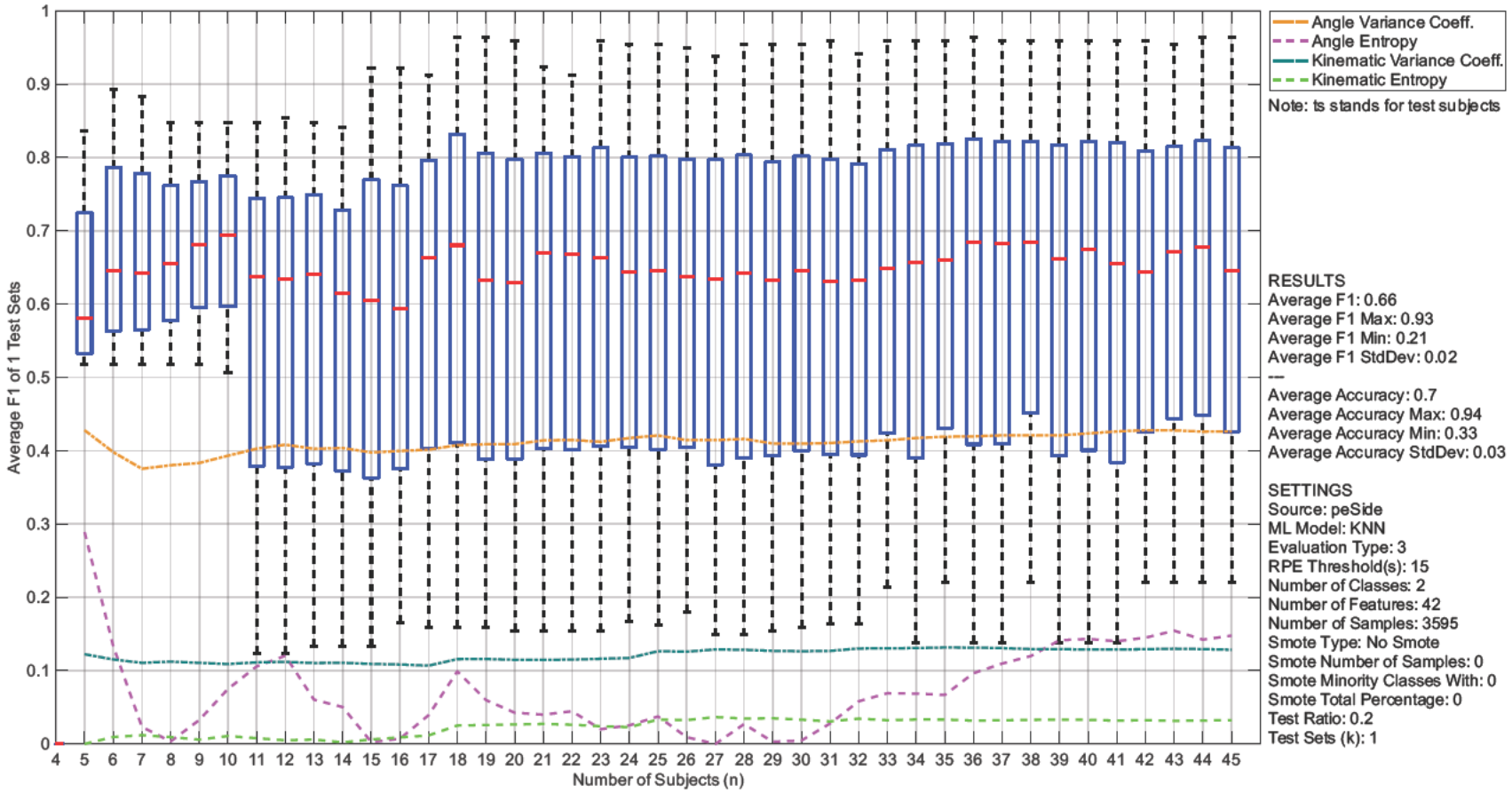


Fig. 5.4.: $_{\text{macro}}F_1$ scores for an incremental number of subjects with *T3-LOSO* evaluation.

5.10.5 n Models with *T2-LNSO* Evaluation

Figure 5.5 shows the ${}_{\text{macro}}\text{F}_1$ scores using *T2-LNSO* evaluation with 5-fold cross-validation. The minimum and maximum ${}_{\text{macro}}\text{F}_1$ scores are almost identical to the median, which is why the boxes are only visible as horizontal straight lines. The performance increases from around 80% at the beginning to 88% at n=16, after which it fluctuates between 85% and 89%.

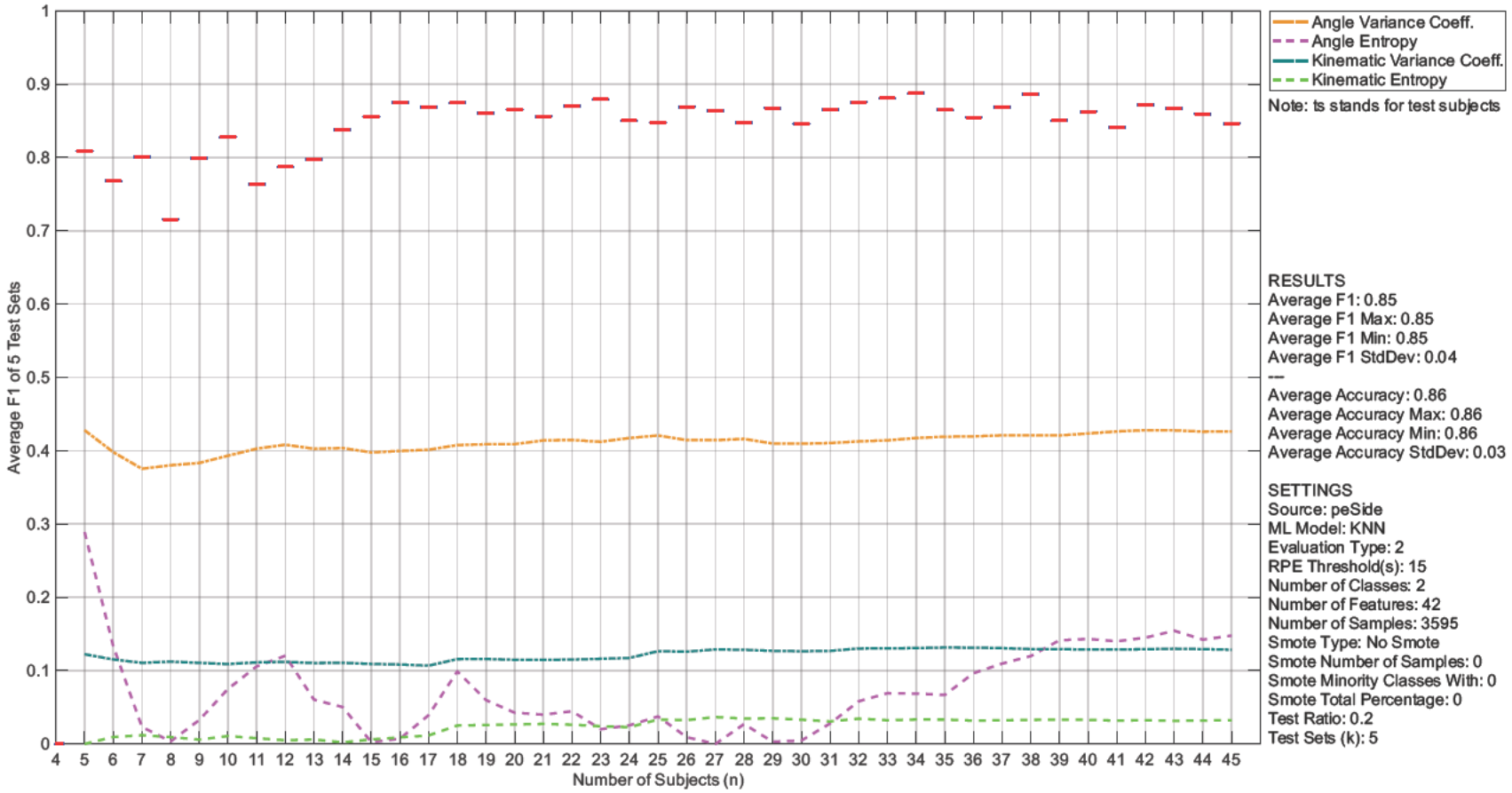


Fig. 5.5.: $_{\text{macro}}F_1$ scores for an incremental number of subjects with *T2-LNSO* evaluation.

6 Discussion

This chapter analyses the results of the case study presented in Chapter 5. By addressing key considerations such as evaluation types, imbalanced data, data augmentation, and model generalisability, this chapter contributes to a deeper understanding of the complexities inherent in exercise fatigue detection with small data.

Section 6.1 presents a comparative analysis of ML models trained on either IMU or PE data. Section 6.2 investigates the generalisability of fatigue detection ML models in regard to different metrics, evaluation types, imbalanced data sets, and oversampling techniques. Section 6.3 discusses statements made in related works about generalisability. Section 6.4 discusses barriers to the development of generalisable exercise fatigue detection ML models and potential solutions.

6.1 Comparison of IMU and PE

The ML models trained on IMU and PE data sets showed similar classification performance (see Table 5.5). The identical class distributions in both data sets facilitated a comparative analysis. However, it is important to note that while identical class distributions help, they do not guarantee that the data sets are comparable in all respects. Other factors such as feature space, data quality, or underlying data transformation processes could still introduce bias.

Performance in Relation to Data Source

While IMU and PE-Front achieved an average accuracy of 0.72 ($_{\text{macro}}F_1$ scores: IMU 0.67 and PE-Front 0.70), the PE-Side models performed better (accuracy 0.77 and $_{\text{macro}}F_1$ score 0.74), probably due to the clearer visibility of joint angle changes from this perspective during squats. These results suggest that PE is a viable alternative to IMU, especially where mobility, privacy, and camera limitations are not critical. PE also offers potential advantages over IMU, including richer information (e.g., joint angles or interaction detection) and faster setup. While it is possible to detect interactions or joint angles using multiple IMUs, such systems often require calibration and more complex setup as the sensors need to be positioned accurately beforehand (see Figure 6.1).

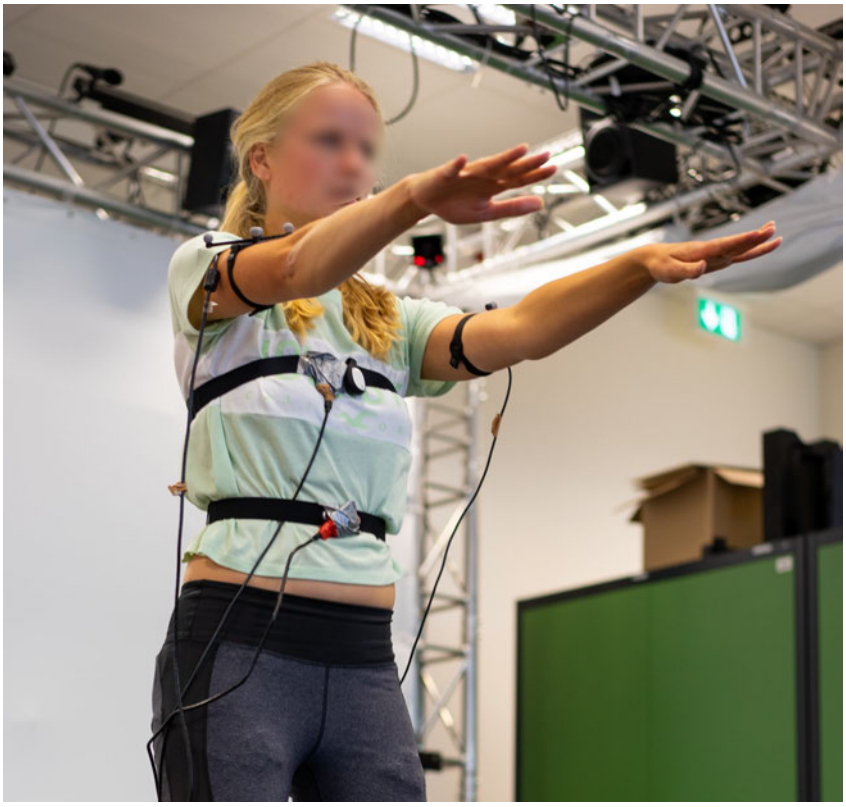


Fig. 6.1.: Preliminary tests with four IMUs placed on the abdomen, sternum and both shoulders, plus two marker-based trackers.

Performance in Relation to Feature Sets

The performance of different feature sets for IMU and PE was investigated.

IMU: In the case of IMU, a model trained with features from both integrated sensors (accelerometer and gyroscope) achieved an average accuracy of 0.72 and ${}_{\text{macro}}\text{F}_1$ score of 0.67 (see Table 5.7). Models trained with gyroscope features alone achieved a slightly higher average accuracy of 0.73, but with a wider range between the minimum and maximum values. This suggests that the combined feature set of accelerometer and gyroscope can provide a more stable performance.

PE: In the case of PE, separate feature sets were examined for the shoulder, hip, and knee. For PE-Side, the shoulder-based feature set showed the highest classification performance with an average accuracy of 0.78 and a ${}_{\text{macro}}\text{F}_1$ score of 0.74 (see Table 5.8). Conversely, for PE-Front, the shoulder and hip-based features were the best, with the hip having a narrower range between minimum and maximum values, resulting in an average accuracy of 0.77 and a ${}_{\text{macro}}\text{F}_1$ score of 0.71 (see Table 5.9).

It is interesting to note that the features of the knee joints did not produce the best results, even though they appear to be the most relevant for squats. As far as fatigue is concerned, the shoulder and hip joints are more supportive for the model, possibly because the movement in these joints changed more with increasing fatigue.

ML models trained with features derived from the joint angles of the shoulder, hip, and knee from the PE side achieved the best overall performance. In contrast to the IMU, a combination of kinematic and angle features resulted in slightly poorer performance, although the range between the minimum and maximum values remained consistent. These results suggest that the side view and joint angles provided the most relevant information for squat fatigue detection.

Conclusion

The comparative analysis suggests that both IMU and PE data can effectively support squat fatigue detection. PE-based models, particularly those using side view data, outperform IMU models in terms of classification accuracy and $_{\text{macro}}F_1$ scores. The difference may be due to the changes in joint angles that are best captured by PE-Side during squats. While this case study focused on squats, the potential applicability of PE to other exercises warrants further investigation.

Another possible avenue for further research could be the fusion of IMU and PE data to potentially improve classification performance or robustness to overcome their respective limitations. However, it is questionable whether this approach is feasible in real world scenarios as fixed cameras contradict the mobility of wearable IMUs.

6.2 Pitfalls of ML with Small Data

The following sections describe the pitfalls of individual-based ML with small data that have been identified based on the literature and the case study. References to appropriate strategies described in the Fatigue Recognition Chain are provided at the end of each pitfall.

6.2.1 Pitfalls based on the Literature

Based on the literature review (see Chapter 2), the following potential pitfalls in ML with small data have been identified.

Inadequate ML with Small Data

ML often performs poorly on small data sets [115]. ML algorithms and traditional data mining processes typically require a large amount of data to train the algorithm-

specific model [226]. ML, especially deep learning, can effectively learn with big data sets. However, it cannot effectively learn with small data sets due to various problems such as overfitting, noise, outliers, and sampling bias, which can render the learned model virtually useless [204]. ML can learn to replicate data bias, and worse, can sometimes amplify the bias [25] – and a small data set is statistically more likely to be biased. Since exercise fatigue detection is based on small data sets, it is an open question whether ML is the right tool at all, especially since it often lacks explainability.

See also 3.6.8 Sample Complexity, Sample Variability, Sample Size, and Sample Bias.

Inadequate Sample Selection / Complexity

Small data sets may not capture the full range of variation in the data. When the group of subjects is highly diverse, training a generalisable model can be challenging. To address this, Kathirgamanathan et al. [191] used clustering to group similar subjects. This allows comparable ML models to be trained for each group, potentially improving homogeneity at the cost of smaller training sets per group. Rather than trying to fix an inhomogeneous group afterwards, it is probably better to define a clear target group or target application and then estimate the complexity of the samples and the amount of data required in advance.

See also 3.1.6 Sample Selection and 3.6.8 Sample Complexity and Sample Variability.

Inadequate Research Design

The intensity of the exercise can influence the onset of fatigue. For example, if the exercise intensity is not appropriate, a timed exercise may end before the subject begins to feel fatigued [102]; in a heterogeneous group this is likely to lead to imbalanced data sets.

See also 3.1 Step 1: Foundational Characteristics.

Inadequate Sensor Selection

Sensor selection and placement can have a considerable impact on model performance [247]. The choice of sensors and their placement should be tailored to the specific exercise being studied.

See also 3.1.5 Sensor Selection.

Inadequate Feature Selection

It is possible to use features that are not relevant, or similar features that do not extend each other. However, the number of features should generally be reduced to the most relevant [281, 308].

See also 3.4.1 Feature Dimensionality, 3.4.2 Feature Extraction, 3.4.3 Feature Transformation, 3.4.4 Feature Selection, and 3.4.6 Feature Normalisation.

Inadequate ML Model Complexity

Some custom ML (deep learning) models proposed in the related works appear to be complex, lack explanatory power, and generalisability, given that they have been tested on a small data sets [75].

See also 3.6.8 Model Complexity.

6.2.2 Pitfalls based on the Case Study

Based on the results of the case study (see Chapter 5), the following potential pitfalls of ML with small data have been identified.

Direct comparisons of ML models trained on different data sets can be challenging. To facilitate a more reliable assessment of generalisation, the models in the case study were trained and tested on similar data sets with comparable subjects and class distributions (see section 5.10.1). This helps to isolate the effects of the ML methods themselves, minimising the influence of data-related factors. It is important

to note that this approach is not without limitations. There may still be subtle differences between the data sets that can affect model performance. Furthermore, the ML models presented in Chapter 5 are mainly based on *k*-NN. However, similar results obtained with other models.

Metrics

Accuracy and $_{\text{macro}}F_1$ scores should be taken with caution, as shown for the *k*-NN and SVM models in Section 5.2. The SVM model achieved a higher accuracy score (0.80 vs 0.77), but its $_{\text{macro}}F_1$ score was lower (0.73 vs 0.77). The use of accuracy alone would not provide a sufficient assessment of the generalisability of the ML model, especially due to the presence of imbalanced classes [355, 204, 25]. 21 of the 95 related works relied only on accuracy as a measure (see Section 2.4.2).

In contrast, the NB model achieved higher $_{\text{macro}}F_1$ scores than accuracy (see Section 5.2). When a model achieves high precision and recall for the minority class, it may result in higher $_{\text{macro}}F_1$ scores compared to accuracy. This is probably due to the class imbalance in the data set, with fewer samples in the fatigue class. While the model may perform well in predicting the majority class, it may struggle with the minority class, leading to misleading performance results. Accuracy in this case does not take into account class imbalances and may not reflect the true performance of the model.

As shown in the confusion matrices in Figure 5.1, the SVM models sometimes failed to detect the fatigue class for some test sets. This highlights the importance of confusion matrices in identifying challenges in predicting specific classes for ML models. To achieve a more comprehensive evaluation, multiple metrics such as confusion matrix, specificity, precision, and recall should be used[1].

Some related works concluded that their trained ML models did not overfit and thus generalised, but this assumption may be inaccurate, especially if details about

[1]Multiple metrics were computed and analysed for all models in this thesis, although accuracy and $_{\text{macro}}F_1$ scores are primarily presented.

the distribution of classes and samples are unknown. In addition, most of the related works that included confusion matrices presented each class as a percentage rather than an absolute number of samples, which can be misleading since a score of, e.g., 83.3% in one class could be based on a total of 6 samples, which is probably not sufficient for generalisation.

See also 3.1.8 Unit of Analysis, 3.6.1 Classification Metrics, 3.6.2 Regression Metrics, and 3.6.6 Cross-Validation.

Oversampling

SMOTE was used as an oversampling technique to generate artificial samples for the minority class(es) to mitigate class imbalances, potentially leading to ML models that generalise better across classes. However, the oversampling results presented in Section 5.7 showed reduced average accuracy and ${}_{\text{macro}}\text{F}_1$ scores; higher oversampling ratios corresponded to even worse performance of the ML models. This was observed for all ML methods.

The poorer model performance may be due to the overall scarcity of data for the minority classes, and the available data may not adequately represent all potential variations. Another potential problem is that SMOTE artificially augments the data across all dimensions of the feature vector, potentially creating samples that do not occur naturally. For example, if only some features vary within the minority class, SMOTE will still generate samples where all features are altered, resulting in artificial samples that do not reflect the original data set. This could introduce additional noise, making it harder for the model to identify patterns reliably.

In addition, it is suspected that the boundary between fatigue and non-fatigue classes in the data is narrow, implying that even minimal changes may be decisive and that SMOTE makes too many changes. A priori feature dimensionality reduction or a more intelligent oversampling algorithm that compares the changes of multiple samples with respect to features of the same class (feature relevance) and only

changes relevant features may be worth exploring (see also Pradipta et al. [279] and Fernández et al. [117]). However, another plausible explanation for the failure of synthetic samples to accurately represent the underlying data could be the presence of complex patterns or non-linear relationships within the minority class.

See also 3.3.3 Data Augmentation and 3.4.5 Feature Augmentation.

Intra- vs Inter-Oversampling

The effects of intra-SMOTE (using data from the same subject) versus inter-SMOTE (using data from different subjects) on model performance was evaluated, as shown in Section 5.7. Intra-SMOTE achieved better classification performance, with an average accuracy of 0.71 and a ${}_{\text{macro}}\text{F}_1$ score of 0.68, compared with 0.70 and 0.64 for inter-SMOTE. A combined approach yielded the lowest performance with an accuracy of 0.62 and a ${}_{\text{macro}}\text{F}_1$ score of 0.58.

It was hypothesised that intra-SMOTE may produce more realistic samples, as they originate from the same individual, whereas inter-SMOTE uses data from multiple subjects and may introduce less realistic synthetic samples. However, due to the issues described in the previous section, this was not investigated further, although it may be worth exploring in future research.

See also 3.6.8 Intra- and Inter-class Variability.

k-Fold Cross-Validation

Due to small data, some subject-based folds may not include all classes due to limited class diversity within individual subjects, which could result in a division by zero when calculating F_1 scores. To achieve a robust evaluation, the partitioning function should ideally generate stratified folds that include all classes and maintain a balanced class distribution (see for example Bustos et al. [58]). Such issues were rarely mentioned in related works, although this is a likely phenomenon when working with small data.

See also 3.6.6 Cross-Validation.

Evaluation Types

The chosen evaluation type affects the performance of the ML models, as shown in Section 5.5. It is therefore important to clearly describe how ML models have been evaluated, which has not always been the case in the related works (see Table B.4 in the Appendix). In the case study, the $_{\text{macro}}F_1$ score was 0.87 for *T2-LNSO*, 0.67 for *T3-LOSO*, and 0.77 for *T4-LMSO*.

T2-LNSO *T2-LNSO* evaluation is not suitable for individual-based ML. Although a model evaluated with *T2-LNSO* is tested with unknown data, it is not evaluated with data from unknown subjects. This means that the trained model has at least some knowledge of each subject in the test set. In the case study, *T2-LNSO* achieved the highest scores with an average accuracy of 0.88 and $_{\text{macro}}F_1$ 0.87.

T3-LOSO *T3-LOSO* evaluation is suitable, but the test ratio is often low, especially the higher the number of subjects in a study (in the related works, a median test ratio of 15% was used. See Section 2.4.2). In the case study, 48 subjects were recruited, which corresponds to a test ratio of 2.17% with *T3-LOSO*. When using *T3-LOSO*, it is essential to create a separate test set for each subject (k-fold) and then compute the average performance across all test sets (folds). The overall performance of an ML model cannot be fully reflected by using only a single subject for evaluation. Some related works did not clearly specify whether the reported results were averaged over all test sets or not.

In the case study, the $_{\text{macro}}F_1$ scores for individual subjects ranged from 0.21 to 0.96. The high variability also led to a lower average $_{\text{macro}}F_1$ score of *T3-LOSO* (0.67) compared to *T4-LMSO* (0.77). When *T3-LOSO* is used, the range of results for individual folds/subjects should also be reported, as high variability may indicate poor generalisability.

T4-LMSO For *T4-LMSO* evaluation, test sets should include a sufficient number of subjects to achieve an appropriate test ratio. As described above for *T3-LOSO*, it is recommended to calculate average metrics across multiple test sets (k-folds) and

to analyse the results of individual test sets for variability. While *T4-LMSO* offers advantages in terms of higher test ratios, it does not fully reflect real world scenarios where predictions are typically made for individual subjects rather than groups.

See also 2.4.2 Evaluation Types, 3.6.7 Evaluation Types, and 6.2.2 Evaluation Types.

Number of Classes

Accuracy and ${}_{\text{macro}}\text{F}_1$ score decreased with more classes, as shown in Section 5.4. The more classes, the less sample data was available for each class. The choice of how many classes to use depends on how the samples are distributed across the classes. With limited samples per class, the number of feasible classes for classification decreases, as there may not be enough samples to adequately train each individual class. In addition, the ability to discriminate between classes plays a critical role. In the case study, it appeared to be a challenge for the ML models to discriminate between multiple classes of fatigue. As shown in Table 5.2), the ${}_{\text{macro}}\text{F}_1$ score decreased as the RPE threshold was shifted downward for the binary model (15→0.83, 14→0.74, 13→0.62, 12→0.60). This threshold shift also changed the distribution of the samples by moving more samples from the majority class to the minority class, but the models still struggled to discriminate the classes. This difficulty may also be due to the chosen hyperparameters, features, and the study design. Regarding the study design: Fatigue may accumulate gradually and increase sharply only towards the end of an exercise, suggesting that there may be only two practically distinguishable classes in the sample data. A different study design may be more appropriate for effective multi-class classification, highlighting the importance of study design to the outcome.

See also 3.6.8 Sample Variability and 3.4.5 Imbalanced Classes and Feature Augmentation Approaches.

Regression

Regression could serve as a potential alternative to classification (see also "The dangers of categorical thinking" by Langhe and Fernbach [221]), bypassing the need for arbitrary thresholds. Furthermore, determining an exact RPE may not be critical for certain applications, as RPE are inherently subjective. For example, a model that predicts values such as 14.3 or 16.1 instead of the actual subjective value of 15 may still be acceptable. However, regression does not solve the problem of imbalanced data and generalisability.

Among the evaluated regression models, as shown in Section 5.9, the Gaussian model achieved the lowest MAE value of 2.57, based on an RPE scale of 6 to 20. However, Figure 5.2 shows notable deviations of the model across all RPE, particularly at the boundaries – below 9 and above 14 – where the number of samples is sparse. Consequently, the ability of the model to reliably predict RPE in these regions may be limited, including its generalisability.

See also 3.2.4 Labelling and 3.6.2 Regression Metrics.

Incremental Number of Subjects

It has been shown that increasing the amount of data can help to improve the generalisability as well as the overall performance of the model [166]. In order to assess the generalisability of the ML models, this section discusses how their performance evolved with increasing n (see Section 5.10). To do this, *k*-NN models were trained and tested with different numbers of subjects on PE-Side data – similar trends were observed for other models. A consistent test rate of about 20% was maintained (see Section 5.10.1). As n increased, new subjects were gradually introduced and the overall data distribution changed slightly; the higher the n, the less change there was. However, complete consistency between data sets was not possible, so direct comparisons of models should be treated with caution due to individual data set differences.

Variability Figure 5.3 illustrates the evaluation with *T4-LMSO*. It shows how the IQR as well as the minimum and maximum range of $_{\text{macro}}F_1$ scores across the 10 test sets gradually decreased as n increased. This convergence resulted in a narrower range of $_{\text{macro}}F_1$ scores as n increased. One factor contributing to this trend was the gradual increase in the number of subjects per test set. For test sets consisting of only a few subjects (less than 8 subjects), the $_{\text{macro}}F_1$ scores showed fluctuations, sometimes exceeding 10 to 25 percentage points. This variability highlights the strong dependence of fatigue detection performance on individual subjects and the distribution of their respective samples. Moreover, some subjects' samples appear to be more distinct, leading to sometimes large variations in model performance when a new subject is added.

Conversely, as the number of subjects in the test sets increased (in this case to at least 8), the consistency of the performance of the trained models improved. However, for *T3-LOSO* evaluation (see Table 5.4), even with a larger sample size (n=45), the range of IQR and min-max performance of the trained ML models remained substantially large. This raises questions about the practical generalisability of a model trained with n=45 when applied to individual predictions. These observations underscore the importance of examining the possible range across k test sets, rather than relying on an accumulated result from multiple test sets, or even from one specific test set (e.g., by cherry-picking the best performing test set). Such details should be documented to improve the reproducibility and transparency of research results.

Overall, the more subjects are used as a test set, the lower the variability of the predictions, as shown in Figure 5.3, because outliers in the test set become less significant. For this reason, the test ratio is important and should not be too small, as is typically the case with *T3-LOSO* evaluation, leading to models that are sensitive to particular subjects. Unfortunately, *T4-LMSO* evaluation is not easily transferable into practice, because usually only the fatigue of one person needs to be predicted, not that of a group of people. The dilemma for individual fatigue detection is that

T3-LOSO evaluation is required, but inevitably has a low test ratio, leading to high variability with small data. It remains an open question whether this variability will decrease with a very large number of subjects (and thus a very large training set), and how large it needs to be.

In contrast, for *T2-LNSO* evaluation (see Table 5.5), the minimum and maximum $_{\text{macro}}F_1$ scores (variability) were almost identical to the median, while performance increased from around 80% at n=5 to 88% at n=16, after which it fluctuated between 85% and 89%. This shows that even small amounts of data are sufficient to achieve relatively high scores with *T2-LNSO* evaluation.

Homogeneity The variance coefficient and entropy were also computed from the raw data for each incremental n (see Section 5.10). These metrics generally increased slightly with increasing n, but were relatively stable without major fluctuations. An exception was the entropy of the joint angles, which fluctuated to some extent up to n=30, until it appeared to stabilise around n=45. Since the variance coefficient of the joint angles remained relatively constant, it is assumed that the noise in the raw data was the reason for the increased entropy.

These metrics can be useful in estimating how homogeneous the subjects were as a group: the more consistent they are for each increment of n, the more likely the subjects are homogeneous. Homogeneous groups can also increase the likelihood that all possible variation can be captured with small data sets. Inhomogeneous groups are more likely to lead to greater variation in model performance, which will then depend on the individual subject(s).

Closely related to homogeneous groups is the study design, which includes sample and exercise selection. The quality of exercise performance can vary from person to person and often deteriorates due to fatigue (see also Appendix O). Individuals may compensate for difficult exercises with avoidance movements. To achieve comparable exercise performance, changes in technique and form should be monitored and corrected if necessary. In general, it is important to make sure that the exercises are

challenging but not too difficult. If subjects can be divided into subgroups, such as slow and fast fatiguing, this may indicate that the target group has not been well specified or that a larger data collection is required. Defining clear limitations is necessary to carefully balance sample variability, exercise complexity and sample size while minimising bias (see also Section 3.6.8).

See also 3.6.3 Visualisation and 3.6.8 Underfitting and Overfitting, Sample Variability, and Sample Size.

ML Tailoring

Overfitting is a common problem with small data sets, since most ML models can memorise most of the data set. Tuning ML models on small data sets should be done carefully: While hyperparameter tuning and feature engineering can improve performance and generalisation through regularisation, they can also be counterproductive if the data is biased or unrepresentative. Excessive tuning on small data can also lead to fragile models that do not overfit and perform well on the training and test data but struggle to generalise to new data. This is particularly likely for models trained in controlled laboratory environments or on unrepresentative test subjects, which may not fully represent real world conditions.

See also 3.4.1 Feature Dimensionality, 3.5.3 ML Training, 3.6.4 Optimisation, and 3.6.8 Model Complexity, Sample Complexity, Sample Size, and Sample Bias.

6.3 Generalisability Myths

This section discusses statements found in the related works about generalisability of trained ML models (see also Table B.7 in the Appendix). The following statements are sorted by frequency, with the most frequently mentioned first.

Myth 1: Cross-validation reduces overfitting and/or assesses generalisability. Cross-validation by itself does not reduce overfitting, but it can help to identify overfitting.

To reduce overfitting, regularisation techniques (such as L1 or L2), dropout (for neural networks), or pruning (in decision trees) are typically used. Cross-validation does not guarantee generalisation to new data distributions. If the training data doesn't adequately represent real world scenarios, cross-validation results can still be misleading. If there is a systematic bias in the data set, cross-validation won't necessarily detect it. While cross-validation can help to identify overfitting, it does not inherently diagnose the cause of overfitting, e.g., data quality and model complexity. Cross-validation can also introduce bias in small data sets due to a low number of test samples, affecting the stability of the performance estimates.

See also 3.6.6 Cross-Validation and 3.6.8 Generalisation.

Myth 2: Leave-one-subject-out evaluation assesses or ensures overfitting and/or generalisation. *T3-LOSO* can assess overfitting and provide an estimate of model generalisation, but it does not prevent or guarantee it. If the model is too complex or overparameterised, it may still overfit. *T3-LOSO* often has a low test ratio and high variance in its estimates, especially with small data sets, which can lead to an unstable measure of generalisability.

See also 3.6.7 Evaluation Types, 3.6.8 Generalisation, and 6.2.2 Evaluation Types.

Myth 3: Feature selection/reduction prevents overfitting and/or improves generalisability. While feature selection can help mitigate overfitting by reducing model complexity and dimensionality, it does not prevent overfitting. The degree to which overfitting is prevented depends on several factors, including model architecture, regularisation techniques, and the size and quality of the training data. Feature selection can improve generalisation by focusing on the most relevant features, but it can also reduce generalisation if the selected features are overly tailored to a small or biased data set, leading to poor performance on unseen data. Furthermore, the impact of feature selection varies between different ML models. For example, linear models (e.g., linear regression) are particularly sensitive to irrelevant or noisy features, making feature selection more important. On the other hand, tree-based

models (e.g., random forests) tend to be more robust due to their inherent ability to rank and ignore less important features through splitting criteria and feature importance measures.

See also 3.4 Feature Engineering and 3.6.8 Generalisation.

Myth 4: The use of ML method XYZ would improve generalisation. Generalisation is influenced by several factors, including the architecture of the model, the size, balance, and quality of the training data, regularisation techniques, data augmentation, and cross-validation. While some ML methods, such as RF and SVMs, have built-in mechanisms to mitigate overfitting, these methods are not infallible. Overfitting can still occur, especially if the training data is small or unrepresentative.

See also 3.5.1 ML Method Selection and 3.6.8 Generalisation.

Myth 5: RF provides robustness to small data sets. RF benefits from a large amount of data to build many different trees, and with small data sets the randomness in bootstrapping can lead to instability or redundant trees that do not improve model performance. Moreover, RF can struggle with small data sets because the DT are constructed by splitting data recursively, which requires sufficient data to avoid overfitting at each split – the similar problem exists with k-fold cross-validation. Therefore, RF performance can be sensitive to hyperparameters, such as the number of trees, maximum depth, minimum samples per split, minimum samples per leaf, maximum features, and splitting criterion. RF can be effective on small data sets if the hyperparameters are carefully tuned.

See also 3.5.1 ML Method Selection and 3.6.8 Generalisation.

Myth 6: Bagging reduces amount of variation in a data set and reduce the amount of overfitting. Bagging creates multiple models by training them on different bootstrapped subsets of the original data set, where each subset is sampled with replacement. This can increase model diversity, especially for high variance models such as DT, because each model is trained on a slightly different data set. By averaging the predictions of multiple models, bagging can reduce overfitting by reducing the

variance of individual predictions and improve overall generalisation performance. However, while bagging does not reduce the variation in the data set itself, it does use variation of multiple models to increase the robustness of the ensemble.

See also 3.6.5 Bootstrapping and 3.6.8 Generalisation.

Myth 7: Splitting data into training and test set prevents overfitting problems and improves generalisability. Splitting the data into training and test sets allows overfitting to be detected and provides a way to assess the generalisability of a model. However, it does not inherently prevent overfitting or improve generalisation.

See also 3.5.2 ML Strategy and 3.6.8 Generalisation.

Myth 8: A minority fatigue class was removed to improve generalisability. Removing a minority fatigue class is unlikely to improve generalisation and may lead to bias and reduced model robustness. Instead, methods for handling imbalanced data should be used to increase the ability of the model to generalise across all classes, including minority cases.

See also 3.4.5 Imbalanced Classes and 3.6.8 Generalisation.

Myth 9: The model can be improved even more by including more features. Adding more features to a model does not necessarily improve its performance and can lead to overfitting, redundancy, and inefficiency. The quality and relevance of features, rather than their quantity, are more critical to improving model performance and generalisability. Feature selection or dimensionality reduction techniques should be used to avoid adding unnecessary complexity.

See also 3.4 Feature Engineering and 3.6.8 Generalisation.

Myth 10: The training set provided to the model was representative of the entire space of interest, so that the trained model had the ability to generalise. Ensuring that a training set fully represents the "entire space of interest" is challenging, especially for small data sets and complex domains such as fatigue. Training data typically covers a limited portion of the possible input space. Generalisation depends on factors beyond the data itself. Model complexity, regularisation techniques, and

the presence of noise play an important role. Systematic biases in the data can also hinder generalisation, even if the data appears to be representative. Distribution shifts can occur if the training data differs from real-world inputs. For example, a model trained on one age group may underperform on another. Changes in the underlying conditions (e.g. athletes changing their training routines) can also lead to distribution shifts. For this reason, it is important to clearly define the target group or application for a trained ML model.

See also 3.6.8 Generalisation.

6.4 Potential Causes and Strategies

This section explores potential causes and strategies for generalisable exercise fatigue detection with small data. Understanding the causes is a critical first step in developing strategies for more generalisable ML models for exercise fatigue detection.

No Big Data

Achieving a balanced data set is a challenge in fatigue research. Unlike other fields, the nature of fatigue limits the collection of large amounts of data from fatigued subjects, even though prolonged training sessions in a fatigued state are impractical. Given the variation between classes, it is challenging to find an optimal balance between recording time and inducing fatigue in most subjects, while ensuring a gradual onset of fatigue. In addition, obtaining fatigue data requires complex and time-consuming studies with real subjects, as opposed to big data settings where data collection, such as logging orders from a web shop, can be done passively. This difference highlights the complexities and limitations associated with obtaining sufficient and representative data for fatigue research.

Imbalances

Imbalances between classes pose an additional challenge for fatigue detection, as there is little data available for the minority classes. In fatigue detection research, there are often not enough samples from the minority class to adequately train the model. For example, in a data set with one million samples and a class imbalance of 10:1, 100000 samples would still be available to train the minority class. However, if the data set contains only 1000 samples with the same class imbalance, 100 samples would be available to train the minority class. The number of samples is further reduced by applying k-fold cross-validation. 15.8% of the related works treated imbalanced data through the use of oversampling or undersampling techniques. 44,2% of the related works did not report if classes were balanced or imbalanced. Details of sample and class distribution were also rarely provided. The following section may explain this.

Powerful ML Tools

Fatigue is a complex concept with different definitions in different scientific fields. To achieve that ML models for fatigue detection are developed and evaluated effectively, collaboration between experts from different disciplines is essential. This interdisciplinary approach can help to gain a comprehensive understanding of both the technical and practical aspects of fatigue detection.

The widespread accessibility of ML methods allows them to be used by people with limited experience or expertise in data science (including ML) [194, 355]. ML in human-centred computing can be performed by researchers (from unrelated disciplines) without knowledge of the underlying data science principles [96, 148, 197]. In addition, the development of automated ML further simplifies the process of using ML by automating many of the tasks involved, making it more accessible to practitioners with less technical expertise [205, 11, 290]. The focus on ML in recent years may also overshadow the importance of the underlying data science

and statistical principles by relying on training, test, and validation sets that do not require specific statistical or mathematical knowledge to identify problems. Moreover, many complex ML systems lack interpretability despite producing highly accurate results [148]. Some ML methods can even fit random data [363]. However, the use of complex ML methods, even for small data sets, can be tempting as it is often accompanied with an improvement in model performance – at the expense of interpretability [75], which is perhaps not prioritised.

Statistical methods, such as linear regression, offer high interpretability but may have lower predictive power compared to ML methods (see Figure 6.2). The choice between these methods depends on the specific problem and desired outcome. If understanding how results are generated is critical, statistical methods may be preferred. If prediction accuracy is the priority, ML may be more appropriate [146].

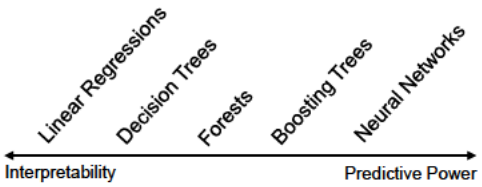


Fig. 6.2.: Example for interpretability vs predictive power by Hassani et al. [148].

Evaluation Methods

47.0% of the related works opted for *T2-LNSO* evaluation. A plausible explanation for the prevalence of *T2-LNSO* may be the default setting of this evaluation type in many ML frameworks and tutorials. Tutorials are often tailored to object, plant, or animal classification, which is an inherently different task from subject-based fatigue detection due to the intersubjectivity of the data – it matters to which subject the samples belong. Some ML frameworks also support leave-one-out evaluation, but additional implementation effort may be required for subject-based cross-validation.

34.0% of the related works used *T3-LOSO* evaluation. However, these studies rarely examined the generalisability of the trained models, nor did they address the challenges posed by low test ratios or report how results varied across different folds. The reasons for this can only be speculated, perhaps it is simply a replication of research methods from other studies, including the lack of further investigation. 6.3% of the related works used *T4-LMSO* evaluation, often for the reason of a higher test ratio.

The Game of Publications

Another contributing factor could be the constraints of publication page limits (e.g., in short papers), so that some researchers may prioritise literature review, results, or discussion over a comprehensive view of the research method and evaluation. Another reason could be that most trained ML models probably do not leave the scientific field and may never be used in real applications, as experiments are mainly conducted for scientific or publication purposes. From 1980 to 2014, the number of all publications increased from one million to more than seven million per year (see Figure 6.3). However, 72% of publications are not even cited once five years after publication [120] and the pressure to publish leads to quantity rather than quality [312]. In addition, studies that fail to find meaningful associations or negative

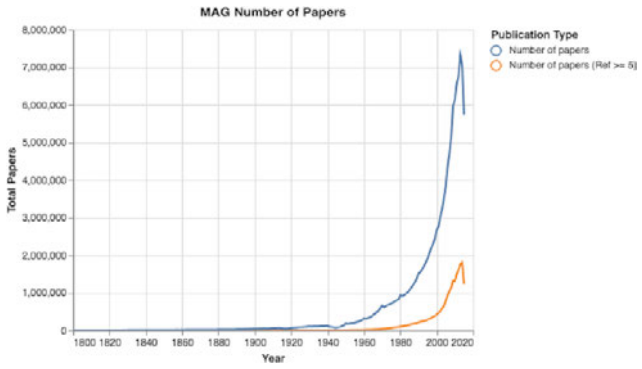


Fig. 6.3.: The total number of publications has surged exponentially over the years. Source: Fire and Guestrin [120].

results may be less likely to be published, leading to a potential overestimation of the effectiveness of fatigue detection methods in the literature.

Reproducibility

Another critical issue (and cause) is the lack of reproducibility in most of the related works; an issue generally highlighted by Peng [271] and Baker [27]. For this reason, dissemination is promoted as a separate step in the Fatigue Recognition Chain (see Section 3.7). Reproducibility is a critical step, as it can lead to more robust and effective ML models by continuously questioning and validating the underlying assumptions and methodologies. The general scientific method is based on falsification which involves actively seeking to disprove existing hypotheses and theories, leading to a more robust understanding of a model. In the context of exercise fatigue detection, this means continually testing and challenging the assumptions underlying the reported models, refining them based on empirical evidence, and being open to revising or rejecting theories that do not withstand scrutiny. According to Yarin [359], there is a need for a hypothesis-driven approach to understand and improve ML tools. A fundamental question is how many publications are needed to demonstrate the ability of ML to accurately detect fatigue patterns.

Vague Methods The research method, algorithms and/or evaluations are often incompletely described, missing essential information. In a literature review, Hussain et al. [162] identified a common problem of missing details in various papers, including essential information such as time and space complexity, latency, classifier used, and clarification on the real-time or offline nature of the proposed techniques, highlighting the importance for authors to provide comprehensive discussions covering these aspects in their approaches. Lin et al. [233] found that the majority of segmentation algorithms reviewed do not explicitly report segmentation accuracy, and they recommended algorithm testing against comprehensive publicly available data sets.

Inaccessible Data Without access to the original data and algorithms used for ML, it is impossible to verify reported results – however, even with all the data and resources available, there remains a high probability that the results will not be reproducible upon review, as highlighted in Peng [271] and Baker [27]. While peer review is a form of quality control, it is not infallible and cannot catch every discrepancy; its effectiveness is also highly dependent on the reviewer [182].

Public databases or benchmarks would allow researchers to test and compare their models, as demonstrated by the online ML platform Kaggle[2], which has coined the "Kaggle Effect", referring to the phenomenon of data scientists and ML practitioners improving and optimising their models by participating in competitions. The Kaggle Effect involves rapid learning, skill building, and cross-pollination of ideas. However, it can also have the opposite effect, where ML models are tailored to a specific data set at the expense of generalisability (as discussed in Section 6.2.2).

Another challenge to the establishment of open data and open sources is the considerable extra effort required, especially to document and provide sufficient metadata to properly describe the semantics of the data [75], which would also require standards.

Lack of Standards Fatigue is a multifaceted phenomenon influenced by various physiological, psychological, and environmental factors (see Chapter 2.2.1). Accurate modelling of fatigue requires capturing these complex interactions, which can be challenging given the limited understanding of fatigue mechanisms and the variability of fatigue experience among individuals. The lack of standardised protocols or benchmarks for evaluating fatigue detection models appears to lead to inconsistencies in performance assessment. Without agreed standards, researchers may use different methods or data sets, making it difficult to compare and validate the effectiveness of different approaches.

[2]Kaggle is an online platform that hosts data science competitions where participants collaborate and compete to create the best solutions.

With standardised protocols in place, a systematic collection of annotated data could be established. This would allow for the accumulation and combination of data sets, as well as the preservation of data over time for trend analysis. This sharing strategy could further promote collaboration, ideas, skills and data quality [198].

6.5 Summary

This chapter presented an analysis of ML for exercise fatigue detection based on the results of the conducted literature review and case study. It focused on the challenges of achieving generalisable ML models. Although the focus of this thesis is on fatigue detection during exercise, many of the challenges described are likely to be applicable to other domains where individual-based ML is used with small data.

Firstly, the performance of models trained on either IMU or PE data was compared. PE-based models achieved better prediction results than IMU models. This may be because they can capture additional information from joint angle changes. While IMUs offer mobility, PE's ease of setup and richer information make it a promising alternative. The analysis also included different feature sets and showed that a combination of accelerometer and gyroscope features gave the best performance for IMU, while shoulder-based side-view features were most effective for PE in detecting squat fatigue.

Secondly, statements found in the related works about the generalisability of trained ML models with small data were discussed.

Thirdly, the possible pitfalls and their causes were explored. Misuse of ML tools: The ease of use of modern ML methods makes it possible to train models that achieve high performance without in-depth expertise. However, not all models that do not overfit to the training data will generalise [363]. Model tailoring: Models tailored to

small data sets from controlled environments, such as laboratories, may not generalise well to real world scenarios. Publication pressure: The emphasis on producing large volumes of research may lead to studies that lack detail and are difficult to replicate, undermining the quality and reliability of the evidence. Research design and subject homogeneity: Building generalisable models requires careful research design and attention to subject homogeneity and sample selection. Small data sets need to capture all possible variation, while still providing sufficient data to avoid class imbalances. Data scarcity: The lack of large, annotated and balanced data sets poses a significant challenge to effective model training and limits the ability of ML models to generalise, leading to the open question of whether ML is the right tool for small data analysis. Inadequate metrics: Accuracy alone is an inadequate metric for evaluating the performance of ML models trained on imbalanced data sets. Additional metrics, such as F_1 score or confusion matrices, are necessary to assess the performance of individual classes and provide a more comprehensive evaluation. Imbalanced data: The presence of many classes combined with small data sets in individual-based ML exacerbates the challenges of minority classes. Details such as sample distribution among the classes are important. Regression could be an alternative to classification to avoid thresholds for classes. Undersampling is usually not feasible, and oversampling can introduce noise. In addition, cross-validation can lead to imbalanced and undersampled folds. Evaluation methods: The choice of evaluation method has a significant impact on model performance. *T2-LNSO* evaluation, which is commonly used as the default method, does not adequately represent unseen test data in individual-based ML. *T3-LOSO* suffers from low test ratios and is highly dependent on the specific subject chosen for testing. *T4-LMSO* mitigates the problem of low test ratios but does not accurately represent how the model would be used in real world scenarios to predict individual fatigue.

The following strategies have been proposed to address these challenges, although they are unlikely to be exhaustive. Future research should build on the findings and explore further strategies.

- Interdisciplinary research: Conduct exercise fatigue detection in collaboration with experts from different disciplines.
- Standardised protocols: Develop consistent methodologies and evaluation benchmarks for exercise fatigue research.
- Rigorous research practices: Prioritise quality, reproducibility, and transparency in research.
- Open mind: Give equal weight to statistics and data science [268, 148]
- Generalisability too: Emphasise generalisability as well as model performance.
- Open science: Share open data, sources, and publications.

7 Conclusion

This chapter concludes this thesis. The following sections reflect on the research aim and questions (Section 7.1), research objectives (Section 7.2), findings (Section 7.3), limitations (Section 7.4), recommendations for future research (Section 7.5), and conclude with a summary (Section 7.6).

7.1 Revisiting the Research Aim and Questions

The aim of this thesis is to identify and address the strategies and pitfalls of sensor-based exercise fatigue detection using ML with small data sets for physical activities such as exercise training in terms of generalisability (see Section 1.1). This section revisits the research questions introduced in Section 1.1.1.

1. How to conduct research on exercise fatigue detection with ML? The Fatigue Recognition Chain is a framework designed to provide a general guide for interdisciplinary researchers to conduct sensor-based exercise fatigue detection (see Chapter 3). The framework covers an incremental process consisting of seven adaptable steps, including fundamental characteristics, raw data collection, data transformation, feature engineering, ML, evaluation, and sharing. Each step considers fatigue, small data, and generalisability where appropriate. To demonstrate the applicability of the framework, a case study of squat fatigue detection based on RPE, IMU and PE data was conducted (see Chapter 4).

2. What are common strategies and pitfalls of ML with small data? Section 6.2 and Section 6.3 discuss several identified pitfalls, each referring to specific strategies described in the Fatigue Recognition Chain in Chapter 3. Typical pitfalls include in-

adequate ML with small data, sample selection, sample complexity, sensor selection, feature selection, model complexity, metrics, oversampling, k-fold cross-validation, evaluation types, number of classes, regression and ML tailoring, as well as various myths, stated by related works, that certain ML techniques ensure generalisation.

3. How do small data, evaluation methods, and augmentation effect ML? Section 6.2.2 describes the various results from Chapter 5 based on the case study conducted for this thesis. In general, the impact on model training can be substantial. Small data sets are likely to be biased or not cover all variations, potentially leading to swingy predictions.

Augmentation techniques (e.g. SMOTE) can address class imbalances in data sets. By creating synthetic data for minority classes, they improve model performance and generalisation. Oversampling also helps to maintain balanced folds in k-fold cross-validation. However, indiscriminate oversampling can introduce noise. It's important to account for the complex patterns and non-linear relationships within the minority class. Oversampling based on specific subjects promises synthetic samples that more closely match the original data, but this needs further research.

The evaluation method has a major impact on generalisability and may be useful for predicting fatigue in individuals (see the following question).

4. How generalisable are ML models trained on small data sets? As discussed in Section 6.2.2, the majority of related works are based on models trained with small data and *T2-LNSO* evaluation, these models are unlikely to generalise properly, because no data from unknown subjects are used for model testing. *T3-LOSO* evaluation is the second most used evaluation type, which is based on a low test ratio and can therefore lead to unstable results depending on the subject to be predicted, especially if the training set is based on a small data set that does not cover all variations. For this reason, *T2-LNSO* models may generalise for certain unknown subjects but not for others. The variation of model predictions for each subject is not usually published, but this should be common practice. *T4-LMSO*

evaluation can be tested with an appropriate test ratio, reducing the effect of an individual subject on the prediction, but such models are best used to predict fatigue in groups of people rather than in individuals.

There is an exception for *T2-LNSO* and *T3-LOSO* evaluation, where models trained on small data can generalise well if the target group is narrowly selected, resulting in a homogeneous target group for model training; such models can generalise to unknown subjects if the unknown subject fits into this homogeneous training group, i.e., the target group is defined in such a way that only certain people are eligible for this model.

7.2 Revisiting the Research Objectives

This section briefly revisits the research objectives introduced in Section 1.1.2.

1. To review the literature on exercise fatigue detection based on sensors and ML. Chapter 2 introduced the key concepts in this thesis: small data in Section 2.1, fatigue detection in Section 2.2, and human activity recognition in Section 2.3. The concepts of generalisation were provided in Section 3.6.8. A literature survey was presented in Section 2.4 including a total of 95 related works.

2. To create a framework for sensor-based fatigue detection research with ML. The Fatigue Recognition Chain framework consists of seven incremental steps and was described in Chapter 3.

3. To conduct a case study with squat exercises by implementing the framework. A case study for fatigue detection based on squat exercises was conducted with a total of n=48 subjects (see Chapter 4).

4. To collect RPE-labelled sensor data from IMU and PE for ML analyses. The case study collected RPE, IMU and PE data (see Chapter 4) which was used to train and analyse various ML models (see Chapter 5).

5. To investigate the ML fatigue predictions with an increasing data set. Section 5.10 presents an analysis of different models with increasing number of subjects.

6. To compare evaluation types and their effect on generalisability. Based on the results presented in Section 5.5, the effects of each evaluation type were discussed in Section 6.2.2.

7. To explore data augmentation techniques to improve generalisability. The prediction results for models trained with oversampled data was presented in Section 5.6 and discussed in Section 6.2.2.

7.3 Findings

Various strategies and pitfalls have been discussed throughout this thesis. This section summarises and integrates the findings of the literature review (Chapter 2), the Fatigue Recognition Chain (Chapter 3), the case study (Chapter 4), its results (Chapter 5), and the discussion (see Chapter 6).

IMU vs PE Both IMU and PE data are effective in supporting the detection of squat fatigue. PE-based models, particularly those using side view data, outperformed IMU models in terms of classification accuracy and $_{\text{macro}}F_1$ scores, probably due to their ability to capture joint additional data such as joint angle changes. Although the case study focused on squats, the potential applicability of PE to other physical activities merits further investigation.

Limited Metrics Accuracy as the sole metric for evaluating ML models can be misleading, especially in the presence of class imbalances. A model may show high accuracy by correctly predicting the majority class, but fail to effectively predict the minority class(es). Instead of relying on accuracy, a range of performance metrics should be considered, such as F_1 score (micro, macro, weighted), confusion matrices, and other means of visualisation. This approach provides a more comprehensive

assessment of model performance, especially when working with small data and imbalanced data sets.

High Variability The performance of ML models can vary considerably depending on the data distribution and individual subjects. Small data sets may not adequately represent the full range of variation, resulting in non-overfitting models that still struggle to generalise to data from new subjects.

Homogeneous Subjects Small data sets carry a high risk of under-representing relevant and potential variation in the data, which can make it difficult to generalise models to unseen data or subjects. To address this, researchers should aim for clear target groups, which requires research designs with specific homogeneous groups to capture as much of the range of variability as possible, even with small data sets.

For effective ML, homogeneity should be considered from the earliest stages of research design. This includes careful sample selection, experimental protocol, ground truth, as well as time and experimental conditions to control the potential variability captured within a small data set during data collection. However, determining the optimal balance between homogeneity, variability, and sample size is challenging, as the resulting model should be representative of its intended application in real world scenarios. The latter should exist and also be used to assess generalisability.

Incremental Training Data A non-overfitting ML model may not generalise beyond the research project. Other factors, such as sample and model complexity, data quality, bias, and representativeness also affect generalisation. Individual-based ML models should be trained on increasing amounts of data to assess their evolving performance. Visualising changes in performance through plots can help identify model stability and weaknesses.

Tailored to the Data Hyperparameters and feature engineering are important factors influencing generalisability. However, if these are not carefully tuned for small data sets, there is a risk that models will be overly tailored to the specific data set, making their predictions less reliable for unseen data.

Evaluation Type Different evaluation methods lead to different results in terms of performance. In individual-based ML, *T2-LNSO* often yields the highest scores, but the generalisability is questionable because the models are trained on data from each subject. For this reason, *T3-LOSO* or *T4-LMSO* should be preferred. However, *T3-LOSO* shows a high degree of variation depending on which subject is chosen as the test person. In addition, the test ratio for *T3-LOSO* is only $1/n$. Averaged results from all subjects (folds) should be used, but the variability of the folds should also be reported. *T4-LMSO* uses a higher test ratio and is less prone to performance variability, but *T4-LMSO* does not reflect the way an application is used: fatigue detection in a single person.

Imbalanced Classes Small data sets often have imbalanced class distributions, especially in fatigue research. ML models may perform well on majority classes but poorly on minority classes, skewing overall performance metrics. In addition, partitioning the data into k-folds for cross-validation can lead to further skewed distributions. It is also important to verify that there are sufficient samples in each fold.

Oversampling Oversampling techniques, such as SMOTE, can be used to balance minority classes by creating artificial data, which can lead to better performance and generalisation of ML models across all classes. In addition, oversampling can help to avoid class imbalances when generating k-folds for training and testing. However, if not applied carefully, oversampling can generate samples that may not accurately reflect realistic data, essentially adding noise, for example, if all dimensions are changed indiscriminately. The presence of complex patterns or non-linear relationships within the minority class should be considered. Rather than using a generic oversampling approach, feature reduction or oversampling methods that modify only relevant features should be considered. Focusing on the relevant features that actually vary within the minority class could provide more realistic synthetic samples.

Another consideration is whether to use data from the same subject (intra-SMOTE) or from multiple subjects (inter-SMOTE). The intra-SMOTE approach has shown better performance, probably because it produces more realistic data that is closer to the original samples from the same individual. In contrast, inter-SMOTE can introduce greater variability that may not accurately represent realistic data (i.e., noise).

Regression vs Classification For tasks, such as predicting RPE in sports, regression may be a suitable alternative. If there are a large number of classes and approximate predictions are sufficient, regression could remove the need for arbitrary thresholds. However, regression does not solve the problems of small data, imbalanced data, or poor generalisation.

Interdisciplinary Collaboration Fatigue is a complex concept with many different definitions. To effectively develop and evaluate ML models for fatigue detection, collaboration between experts from different disciplines is essential. This interdisciplinary approach can help to gain a comprehensive understanding of both the technical and practical aspects of fatigue detection.

Open Science and Benchmarks Where possible, data sets, algorithms, and publications should be made openly available to facilitate testing and comparison of ML models across studies, thereby promoting reproducibility and open science. A description of the semantics of the shared data and algorithms must also be included. Creating and contributing to shared databases can advance the field by encouraging collaboration and knowledge sharing, as well as the accumulation of data over time. This requires ethical approval and participant consent at the start of the study to address privacy concerns.

Rigorous Research Method Research should provide detailed information about the method. This transparency is crucial for reproducibility and meaningful comparisons between studies. The following questions briefly summarise key elements of this thesis that should be considered in similar studies:

- Homogeneity: What are the characteristics of the subjects? What is sample selection and complexity? Are there biases?
- Representativeness: What is the experimental protocol? What is the time and space of the experiments? Do the experiment and subjects reflect the intended application?
- Data Collection: What sensors are used? Where are they located? What are the sensor settings? What are the sensor characteristics? What is the ground truth?
- Data Transformation: What data has been transformed and how? What are the parameters? What thresholds are used? How is the data segmented? Does data transformation introduce noise or bias? Does the transformation alter the distribution? Is the order of the transformation steps important?
- Algorithms: What algorithms are used? What are the parameters? What are the characteristics of the input data? What is the intention of each algorithm used?
- Distribution: How many samples per class or fold exist?
- Augmentation: What techniques are used, including parameters? How many samples are generated (per class)? Are the generated samples based on one or more subjects, one or more features?
- Validation: Has the data been validated? Are constraints or visualisation techniques used? Can the validation be automated?
- Features: What features are used? How many (dimensionality)? Which features are relevant? Are the features comparable, weighted, normalised?
- ML: What ML methods are used? Is the model complexity appropriate? Are hyperparameters adjusted for small data (e.g., regularisation, fold size)? What settings are used for the hyperparameters?

- Evaluation: What metrics are used? Are multiple metrics used? What types of evaluation are used? How is the data partitioned in terms of subjects? How well are individual classes predicted?
- Generalisability: Can the results or the trained model(s) be transferred to real applications? Is the model being evaluated in a real application? How does performance evolve with increasing number of subjects?
- Sharing: How can the data, algorithms, and findings be made openly available?

Standards Standardisation can help to compare results between studies and improve the reproducibility of results. Standardised protocols should therefore be developed and integrated into the Fatigue Recognition Chain.

7.4 Limitations

This work is not without limitations, stemming from a combination of trade-offs, practical constraints, and unforeseen challenges.

Open Data The case study aimed to respect privacy concerns. For this reason, consent from subjects to publish their data was not considered at the outset, and it was not feasible to obtain it retrospectively from all 48 subjects, although open data is promoted in this thesis.

Small Data The case study was conducted with a relatively small number of subject (n=48), which may limit the depth of insight gained from the ML models. Further data collection could provide a broader understanding of how model performance and generalisability evolve across different contexts and subjects. It is likely that the current data set does not fully capture all relevant variation, potentially limiting the analysis.

Synthetic Data This thesis started with a literature review and the case study on fatigue detection – the question of generalisability arose after analysing the results of the case study. Instead of data collection, a purely synthetic data set with controlled

characteristics would be an alternative approach to systematically explore how ML models perform and generalise under different scenarios and conditions (see also [363]).

Laboratory Environment The case study was conducted in a laboratory environment under ideal conditions, which is likely to have influenced the subjects' behaviour and data collection. As noted by Morris et al. [255], a laboratory environment that does not simulate a gym may produce different data than a real world environment. Moreover, the experiments were conducted under special circumstances during the COVID-19 pandemic. Nevertheless, the controlled laboratory environment helped to reduce variables and minimise external factors, allowing for a focused study of generalisability.

Single Exercise The case study was limited to the analysis of squats. Investigating a wider range of exercises could provide valuable comparative insights and demonstrate the transferability of the findings to other exercises. Squats primarily target the lower body muscles and do not induce the same level of overall fatigue as a more cardio-intensive exercise. This, combined with the gradual build-up and rapid intensification of fatigue during squats, is likely to have contributed to the class imbalances. Collecting data from a lighter, but longer, cardio exercise could potentially provide more balanced data due to the slower progression of fatigue.

Homogenous Subjects The recruited subjects of young healthy students may not be as homogeneous as intended. The requirement for the subjects was that they exercised sporadically – there was concern that additional requirements might discourage potential (voluntary) subjects. In retrospect, the requirement was too unspecific. In the case of squats, for example, it is important what kind of sport a subject does on a regular basis. Presumably, experiments with non-athletes are more likely to result in greater data variability than a group of professional athletes in a particular sport. Another limitation is the imbalance in the ratio of male to female

subjects (2:1), which may affect the ability of the trained ML models to accurately detect fatigue levels in female subjects.

Personalised ML *T1-SOLO* evaluation was excluded from this research because it was assumed that personalised ML would generalise better than ML models trained on multiple subjects due to the inherent lower variability in the data. Kathirgamanathan et al. [191] and Dimmick et al. [102] showed that personalised ML models performed better, however, personalised ML models may still face challenges in generalisation due to intra-subject variability and should be further investigated.

Ground Truth Although RPE is widely used in sports science, it has limitations as ground truth. Non-athletes may struggle to accurately assess their RPE due to unfamiliarity with the scale, which starts at 6 instead of 0. This may affect the reliability of RPE as a measure of fatigue. Additional ground truth measures, such as blood samples, could have provided additional verification. However, preliminary tests showed that these methods were not practical for the case study due to the cost and frequency of sampling.

Limited Sensors The choice of sensors for fatigue detection was limited to IMUs and PE, which does not capture all the factors relevant to fatigue. This suggests that the current approach provides an estimate rather than a definitive assessment of fatigue. However, this limitation is intentional, as sensor-based, unobtrusive fatigue detection may be valuable in certain applications, such as unsupervised home exercise or rehabilitation scenarios.

Limited Feature Engineering and Hyperparameters Another limitation of the case study is that ML was limited to statistical features based on features commonly used in related work. Although feature engineering and hyperparameter tuning (regularisation) can improve model performance and generalisability, these strategies were kept to a minimum in the case study to reduce the number of variables. Further investigation in combination with models trained with different numbers of subjects may provide additional insights.

Limited ML Models While the case study used several ML methods that are commonly used in the related works, other ML methods exist. For example, recurrent neural networks and temporal convolutional neural networks are specifically designed for sequential data and could potentially provide better performance when applied to time series data. However, in the case study, a repetition was used as the unit of analysis, allowing comparisons of repetitions regardless of time. Another approach could be to develop custom ML methods for the small data set [154], but verifying the generalisability of these models remains a challenge. Conservative generalisation could also be a viable approach, as it prioritises reliability and generalisation [204]. Another possible approach could be transfer learning [204].

Limited Oversampling The case study examined SMOTE for balancing minority classes, as it is a widely used oversampling technique in the related works. Further research is needed to evaluate the performance of SMOTE when applied selectively to relevant features. In addition to SMOTE, other oversampling techniques could be applied, such as virtual sample generation, which was not applied in the related works, but has been used in other studies on small data [347] and may be worth investigating.

Basic k-fold Partitioning The custom function for partitioning subject data into k-folds was rather basic and could be improved to produce more balanced, stratified folds by evenly distributing the data in terms of subject and class distribution.

Explainability, Fairness, and Privacy ML models should meet the requirements of explainability, fairness, and privacy [75], which are beyond the scope of this thesis.

Fatigue Recognition Chain Evaluation The Fatigue Recognition Chain, although helpful in the implementation of the case study, was only tested by the author who gave it a firm thumbs up – independent testing and evaluation could identify weaknesses and gaps. In addition, the case study focused exclusively on supervised ML. Investigating the applicability of the framework to other approaches, such as deep learning, would be a valuable next step.

Grounded Theory The method used to analyse the related works on small data, imbalanced data, variability, and generalisability was based on categorising and clustering the found statements. More sophisticated techniques, such such as those of grounded theory [86], would probably provide further insights. In addition, several of the related works were rather vague in describing their research method, sometimes requiring indirect derivation of the samples collected, class distribution, and data imbalances, as well as the used evaluation type(s), hyperparameters, data augmentation, and ground truth.

7.5 Recommendations for Future Research

The following steps are proposed to advance research in the field of exercise fatigue detection with individual-based ML:

Open Data, Source, and Science Openly accessible data, sources, and publications for specific fatigue exercise tasks is a crucial first step (see for example [360, 190, 183]). It would enable the reproducibility of published results. It would also allow comparisons between new and existing solutions by testing them under identical conditions [75]. In addition, by storing data over time, it is possible to track trends and patterns, and the longer the record, the greater the ability to build models and have confidence in the conclusions drawn [198].

Comparability can lead to the development of more robust models and a more rapid accumulation of knowledge. Over time, the cumulative value of such infrastructures can increase as data become more readily and widely available. Such a sharing strategy is also more likely to stimulate new interdisciplinary collaborations between researchers and teams, and to foster improved skills through access to new types of data. Data sharing and the adoption of infrastructure standards, protocols and policies can improve data quality and enable third-party verification of data and research. The economies of scale of such infrastructure can also bring financial benefits [198].

In addition, establishing common benchmarks could encourage scientists to share their results and test different approaches. This practice, as well as transfer learning [245], is already well established in other ML fields such as image recognition[1] and large language models[2].

Standardised Frameworks and Protocols The development of standardised frameworks, protocols and semantics for ML in exercise fatigue detection would further streamline research efforts by improving comparability and communication. This would promote consistency and facilitate collaboration between research teams.

Realistic Conditions If research begins in the laboratory, it should not end there. More ML models should be tested in real settings, beyond controlled experimental conditions. Evaluating such models could provide a better understanding of the generalisability and amount of the required training data. The subjects for training the ML models should reflect the target group as closely as possible.

AutoML Automated ML (AutoML) is increasingly being used for ML tasks such as preprocessing, data analysis, algorithm selection, or hyperparameter tuning (see [205, 11, 290]). AutoML could be an additional tool to support scientists and practitioners to apply best ML practices and avoid common pitfalls.

Personalised ML Models Compared to models trained on data from multiple individuals, research in personalised ML, which takes into account individual characteristics [63, 237], is under-represented and tends to perform better [191, 102, 331], as the data is probably more homogeneous (see also *T1-SOLO* evaluation).

Adaptive ML Models ML models that continue to learn and adapt after initial training could be a next step [146]. If initially trained on small data sets, these models could accumulate and integrate more data over time, mitigating the small data problem. Sharing accumulated data between deployed models can further improve their accuracy and generalisability. Continuous learning would also allow ML models to adapt to the specific characteristics of individual users and changing

[1] `https://image-net.org`
[2] `https://www.swebench.com/`

conditions, including factors such as demographics, environmental factors, training routines, and sensor variations.

Compounded ML Models Another approach could be to develop compounded models that combine several specialised ML models, each analysing specific fatigue factors. This approach may more accurately reflect the multifaceted nature of fatigue. The individual results could then be aggregated into a comprehensive fatigue prediction.

Subject-based Oversampling Oversampling algorithms that augment individual-based (i.e., human-centred) data should be further explored. These algorithms should generate realistic synthetic samples by carefully modifying only relevant features and preserving the original data distribution. Research should focus on class imbalances in small data sets through different oversampling and feature selection techniques to identify the most effective approaches to improve model generalisability.

Furthermore, the optimal level of data augmentation and feature augmentation for specific small data sets needs to be investigated. Determining the appropriate level of augmentation can help to maximise its benefits while minimising potential drawbacks.

Advanced Sensors Advances in physics, electronics, and other fundamental fields are leading to the development of novel sensors and devices that provide more efficient signal patterns for detecting human activity [39]. The development of unobtrusive wearable fatigue sensors capable of measuring biosignals in real time would allow a wider range of fatigue factors to be included, potentially leading to more comprehensive fatigue detection models.

Segmentation The motion segmentation in this thesis is based on a semi-automatic process consisting of an off-line analysis using the coordinates from pose estimation, resulting in a perfect segmentation without erroneous segments (see Section 4.3.2).

However, further research should investigate how to train models based on less robust segmentation techniques (see also Lin et al. [233]).

Interdisciplinary Collaboration Collaboration between experts in fields, such as healthcare, sport, psychology, and computer science is essential to develop more holistic fatigue detection models. Such interdisciplinary research could lead to a deeper understanding of the complex factors associated with fatigue.

7.6 Research Summary

The field of exercise fatigue detection faces two major challenges: The multifaceted nature of fatigue and the scarcity of large, high-quality, annotated data sets. Human-centred experiments are often time-consuming and require controlled environments. The diverse factors of fatigue lead to countless approaches to fatigue detection, while the scarcity of data makes it difficult to compare ML models. Overcoming these challenges requires a comprehensive approach.

For this reason, the Fatigue Recognition Chain framework was introduced in Chapter 3 as a guide for conducting interdisciplinary research on exercise fatigue detection with ML and sensor data. The implementation of the framework was demonstrated in a case study with 48 subjects for squat exercises in Chapter 4. Supervised ML was utilised to train different ML models with data from RPE, IMU, and PE. The results were presented in Chapter 5. A comparison between PE and the commonly used IMU showed that an ML model trained on PE data can outperform a model trained on IMU data, suggesting that PE may be suitable for real time fatigue detection.

The key findings of this thesis are based on a literature survey and a case study. They were discussed in Chapter 6. The literature highlights common pitfalls for ML such as inadequate small data, sample selection, sample complexity, sensor selection, feature selection, and model complexity. Building on this, this thesis provided an

in-depth analysis of the pitfalls that occur in individual-based ML with small data sets. Strategies to avoid the pitfalls were presented. These strategies depend on the particular step in the Fatigue Recognition Chain. Therefore, multiple strategies are often needed – depending on the step – to improve the overall generalisability of trained ML models.

A common misconception is that generalisability is ensured if an ML model avoids underfitting and overfitting. However, generalisability requires consideration of many parameters, as discussed in this thesis, such as model complexity, hyperparameters, sample variability, sample complexity, sample size, sample bias, class distribution, feature selection, and representativeness and homogeneity of the collected data.

These parameters must be clearly defined in advance as they determine the total number of variables and therefore the amount of data required. As data in fatigue research is usually small, the number of variables should be minimised, for example, by restricting the people in the target group and by limiting the fatigue-inducing exercise through explicit rules and standards.

The analysis of ML models trained on small data sets requires several metrics, for example, accuracy, F_1 score, and confusion matrices. In addition, the performance of the model should be thoroughly analysed for each class in order to comprehensively cover its predictive capabilities and to identify its weaknesses. In addition to performance, data distribution should be considered – a ML model may achieve 90% accuracy in predicting a particular class, but if only 10 samples were available for that class, the generalisability of the model for that particular class is questionable.

Such imbalanced data can be avoided in advance by a study design that ensures equal samples for each class, e.g., with time-based training exercises and different training weights to label each class. However, this is not always possible, especially as the fatigue class is usually the minority class. A general strategy is to keep the number of classes and (stratified) k-folds to a minimum, so that more samples can be allocated to each class or fold.

Another strategy is oversampling to overcome problems of imbalanced classes and k-folds. However, common techniques such as SMOTE do not take into account the relevance of features and can add more noise to the data set. Oversampling should generate samples that are as close as possible to the existing samples, otherwise they will have too much influence in a small data set.

As the onset of fatigue can cause a shift in the collected samples, poorly chosen fatigue class thresholds can confuse an ML model during training. Class thresholds should also be chosen carefully as they have a large impact on the class distribution and model prediction. Regression may be used to avoid the problem of defining class thresholds.

Tailoring ML models to a small data set should be kept to a minimum: this may produce high prediction results, but is likely to make the models more sensitive to unseen data – even if the trained models do not overfit – because the small training set is unlikely to cover all possible variations. For this reason, the predictions of ML models should be analysed with increasing numbers of subjects (data) to examine how performance evolves, which can potentially be used as an indicator of data saturation.

A crucial choice is the evaluation type, which determines the performance of an ML model and its generalisability. *T2-LNSO* evaluation yields high prediction results, but is unlikely to be generalisable as no unknown subjects are used for testing. Predictions from models trained with *T3-LOSO* evaluation can vary widely from subject to subject due to low test ratios. For this reason, the performance of the model should be analysed for each subject to determine how sensitive the model is to different subjects. The aim should be to reduce the sensitivity of the model, possibly at the expense of accuracy, in order to improve its overall generalisability. *T4-LMSO* evaluation can be used with appropriate test ratios, but should be used to predict fatigue in groups rather than individuals.

Based on the findings, future recommendations were proposed, including the exploration of more supportive, adaptive, and personalised ML; the investigation of

the generalisability of ML models with imperfect segmentation; the development of advanced wearable sensors that can detect biomarkers in real time; the promotion of standards, open data, open science, and interdisciplinary collaboration to ensure reproducibility and comparability of research results.

Fin.

Fin stands for Finally.

Bibliography

[1] Eswar Adapa, Anish C Turlapaty, and Surya Naidu. "Fatigue Classification and Onset estimation using Surface EMG Signals during Strength Training". In: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Oct. 2023. DOI: 10.1109/apsipaasc58517.2023.10317229.

[2] Jesus S. Aguilar-Ruiz and Marcin Michalak. "Multiclass Classification Performance Curve". In: *IEEE Access* 10 (2022), pp. 68915–68921. DOI: 10.1109/access.2022.3186444.

[3] Jordi Aguiló, Pau Ferrer-Salvans, Antonio García-Rozo, et al. "Project ES3: attempting to quantify and measure the level of stress". In: *Revista de neurologia* 61.9 (Nov. 2015), pp. 405–415.

[4] Andrés Aguirre, Maria J. Pinto, Carlos A. Cifuentes, et al. "Machine Learning Approach for Fatigue Estimation in Sit-to-Stand Exercise". In: *Sensors* 21.15 (July 2021), p. 5006. DOI: 10.3390/s21155006.

[5] Md Atiqur Rahman Ahad, Masud Ahmed, Anindya Das Antar, Yasushi Makihara, and Yasushi Yagi. "Action recognition using kinematics posture feature on 3D skeleton joint locations". In: *Pattern Recognition Letters* 145 (May 2021), pp. 216–224. DOI: 10.1016/j.patrec.2021.02.013.

[6] Barbara E. Ainsworth, William L. Haskell, Stephan D. Herrmann, et al. "2011 Compendium of Physical Activities: A Second Update of Codes and MET Values". In: *Medicine & Science in Sports & Exercise* 43.8 (Aug. 2011), pp. 1575–1581. DOI: 10.1249/mss.0b013e31821ece12.

[7] Havva Aktaş and Cem Aslan. "The examination of relationship between body composition and velocıty on amateur soccer players. Çanakkale Onsekiz Mart University Journal of Sport Sciences". In: 1.1 (Dec. 2018), pp. 17–25.

[8] Folami Alamudun, Jongyoon Choi, Hira Khan, Beena Ahmed, and Ricardo Gutierrez-Osuna. "Removal of Subject-Dependent and Activity-Dependent Variation in Physiological Measures of Stress". In: *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 2012. DOI: 10.4108/icst.pervasivehealth.2012.248722.

[9] Justin Amadeus Albert and Bert Arnrich. "A computer vision approach to continuously monitor fatigue during resistance training". In: *Biomedical Signal Processing and Control* 89 (Mar. 2024), p. 105701. DOI: 10.1016/j.bspc.2023.105701.

[10] Justin Amadeus Albert, Arne Herdick, Clemens Markus Brahms, Urs Granacher, and Bert Arnrich. "PERSIST: A Multimodal Dataset for the Prediction of Perceived Exertion during Resistance Training". In: *Data* 8.1 (Dec. 2022), p. 9. DOI: 10.3390/data8010009.

[11]Nadeen H. Alboueishi, Tarek A. M. Nagem, and Esra A. Abdelnabi. “Human Activity Recognition Using AutoML Approach”. In: *2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*. IEEE, May 2024. DOI: 10.1109/mi-sta61267.2024.10599699.

[12]Ethem Alpaydin. *Introduction to machine learning*. Third edition. Adaptive computation and machine learning series. Includes bibliographical references and index. - Description based on PDF viewed 12/23/2015. Cambridge, Massachusetts: MIT Press, 2014. 1640 pp.

[13]Ethem Alpaydın. *Introduction to machine learning*. Fourth edition. Adaptive computation and machine learning. Description based on publisher supplied metadata and other sources. Cambridge, Massachusetts: The MIT Press, 2020. 1691 pp.

[14]Sina Ameli, Fazel Naghdy, David Stirling, Golshah Naghdy, and Morteza Aghmesheh. “Quantitative and non-invasive measurement of exercise-induced fatigue”. In: *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology* 233.1 (June 2018), pp. 34–45. DOI: 10.1177/1754337118775548.

[15]Wim Ament and Gijsbertus J. Verkerke. “Exercise and Fatigue”. In: *Sports Medicine* 39.5 (2009), pp. 389–422. DOI: 10.2165/00007256-200939050-00005.

[16]Adam J. Amorese and Alice S. Ryan. “Home-Based Tele-Exercise in Musculoskeletal Conditions and Chronic Disease: A Literature Review”. In: *Frontiers in Rehabilitation Sciences* 3 (Feb. 2022). DOI: 10.3389/fresc.2022.811465.

[17]Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed. “Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13.5 (Sept. 2016), pp. 971–989. DOI: 10.1109/tcbb.2015.2478454.

[18]A. Angeli, M. Minetto, A. Dovio, and P. Paccotti. “The overtraining syndrome in athletes: A stress-related disorder”. In: *Journal of Endocrinological Investigation* 27.6 (June 2004), pp. 603–612. DOI: 10.1007/bf03347487.

[19]Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. “A Public Domain Dataset for Human ActivityRecognition Using Smartphones. ESANN”. In: *ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligenceand Machine Learning* (Apr. 2013). Ed. by Michel Verleysen. Literaturangaben, pp. 437–442.

[20]Maxwell Fordjour Antwi-Afari, Shahnawaz Anwer, Waleed Umer, et al. “Machine learning-based identification and classification of physical fatigue levels: A novel method based on a wearable insole device”. In: *International Journal of Industrial Ergonomics* 93 (Jan. 2023), p. 103404. DOI: 10.1016/j.ergon.2022.103404.

[21]Shahnawaz Anwer, Heng Li, Waleed Umer, et al. “Identification and Classification of Physical Fatigue in Construction Workers Using Linear and Nonlinear Heart Rate Variability Measurements”. In: *Journal of Construction Engineering and Management* 149.7 (July 2023). DOI: 10.1061/jcemd4.coeng-13100.

[22] Ashrant Aryal, Ali Ghahramani, and Burcin Becerik-Gerber. “Monitoring fatigue in construction workers using physiological measurements”. In: *Automation in Construction* 82 (Oct. 2017), pp. 154–165. DOI: 10.1016/j.autcon.2017.03.003.

[23] Adriana Arza, Jorge Mario Garzón-Rey, Jesús Lázaro, et al. “Measuring acute stress response through physiological signals: towards a quantitative assessment of stress”. In: *Medical & Biological Engineering & Computing* 57.1 (Aug. 2018), pp. 271–287. DOI: 10.1007/s11517-018-1879-z.

[24] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. “Overview and comparative study of dimensionality reduction techniques for high dimensional data”. In: *Information Fusion* 59 (July 2020), pp. 44–58. DOI: 10.1016/j.inffus.2020.01.005.

[25] Ricardo Baeza-Yates. “The Limitations of Data, Machine Learning and Us”. In: *Companion of the 2024 International Conference on Management of Data*. SIGMOD/PODS ’24. ACM, June 2024. DOI: 10.1145/3626246.3656000.

[26] Amir Baghdadi, Fadel M. Megahed, Ehsan T. Esfahani, and Lora A. Cavuoto. “A machine learning approach to detect changes in gait parameters following a fatiguing occupational task”. In: *Ergonomics* 61.8 (Mar. 2018), pp. 1116–1129. DOI: 10.1080/00140139.2018.1442936.

[27] Monya Baker. “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604 (May 2016), pp. 452–454. DOI: 10.1038/533452a.

[28] Konstantinos Balaskas and Kostas Siozios. “Fatigue Detection Using Deep Long Short-Term Memory Autoencoders”. In: *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, July 2021. DOI: 10.1109/mocast52088.2021.9493378.

[29] Derek Ball. “Metabolic and endocrine response to exercise: sympathoadrenal integration with skeletal muscle”. In: *Journal of Endocrinology* 224.2 (Nov. 2014), R79–R95. DOI: 10.1530/joe-14-0408.

[30] Roy F. Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. “Ego depletion: Is the active self a limited resource?” In: *Journal of Personality and Social Psychology* 74.5 (1998), pp. 1252–1265. DOI: 10.1037/0022-3514.74.5.1252.

[31] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, et al. *BlazePose: On-device Real-time Body Pose tracking*. 2020. DOI: 10.48550/ARXIV.2006.10204.

[32] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. “Vision-based human activity recognition: a survey”. In: *Multimedia Tools and Applications* 79.41-42 (Aug. 2020), pp. 30509–30555. DOI: 10.1007/s11042-020-09004-3.

[33] Tim Op De Beéck, Wannes Meert, Kurt Schütte, Benedicte Vanwanseele, and Jesse Davis. “Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, July 2018. DOI: 10.1145/3219819.3219864.

[34]Martin Behrens, Martin Gube, Helmi Chaabene, et al. “Fatigue and Human Performance: An Updated Framework”. In: *Sports Medicine* 53.1 (Oct. 2022), pp. 7–31. DOI: 10.1007/s40279-022-01748-2.

[35]Richard E. Bellman. *Adaptive Control Processes. A Guided Tour*. Princeton Legacy Library. Princeton, NJ: Princeton University Press, 2016. 1 p.

[36]Esther I. Bernhofer. “Investigating the concept of rest for research and practice”. In: *Journal of Advanced Nursing* 72.5 (Feb. 2016), pp. 1012–1022. DOI: 10.1111/jan.12910.

[37]Daniel Berrar. “Cross-Validation”. In: *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2019, pp. 542–545. DOI: 10.1016/b978-0-12-809633-8.20349-x.

[38]Antonio Bevilacqua, Bingquan Huang, Rob Argent, Brian Caulfield, and Tahar Kechadi. “Automatic classification of knee rehabilitation exercises using a single inertial sensor: A case study”. In: *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, Mar. 2018. DOI: 10.1109/bsn.2018.8329649.

[39]Sizhen Bian, Mengxi Liu, Bo Zhou, and Paul Lukowicz. “The State-of-the-Art Sensing Techniques in Human Activity Recognition: A Survey”. In: *Sensors* 22.12 (June 2022), p. 4596. DOI: 10.3390/s22124596.

[40]Valentina Bianchi, Marco Bassoli, Gianfranco Lombardo, et al. “IoT Wearable Sensor and Deep Learning: An Integrated Approach for Personalized Human Activity Recognition in a Smart Home Environment”. In: *IEEE Internet of Things Journal* 6.5 (Oct. 2019), pp. 8553–8562. DOI: 10.1109/jiot.2019.2920283.

[41]Gürkan Bilgin, İ. Ethem Hindistan, Y. Gül Özkaya, et al. “Determination of Fatigue Following Maximal Loaded Treadmill Exercise by Using Wavelet Packet Transform Analysis and MLPNN from MMG-EMG Data Combinations”. In: *Journal of Medical Systems* 39.10 (Aug. 2015). DOI: 10.1007/s10916-015-0304-5.

[42]George E. Billman. “Homeostasis: The Underappreciated and Far Too Often Ignored Central Organizing Principle of Physiology”. In: *Frontiers in Physiology* 11 (Mar. 2020). DOI: 10.3389/fphys.2020.00200.

[43]Ruel Billones, Josephine K. Liwang, Kierra Butler, Letitia Graves, and Leorey N. Saligan. “Dissecting the fatigue experience: A scoping review of fatigue definitions, dimensions, and measures in non-oncologic medical conditions”. In: *Brain, Behavior, & Immunity - Health* 15 (Aug. 2021), p. 100266. DOI: 10.1016/j.bbih.2021.100266.

[44]Attila Biró, Antonio Ignacio Cuesta-Vargas, and László Szilágyi. “AI-Assisted Fatigue and Stamina Control for Performance Sports on IMU-Generated Multivariate Times Series Datasets”. In: *Sensors* 24.1 (Dec. 2023), p. 132. DOI: 10.3390/s24010132.

[45]Attila Biró, Sándor Miklós Szilágyi, László Szilágyi, Jaime Martín-Martín, and Antonio Ignacio Cuesta-Vargas. “Machine Learning on Prediction of Relative Physical Activity Intensity Using Medical Radar Sensor and 3D Accelerometer”. In: *Sensors* 23.7 (Mar. 2023), p. 3595. DOI: 10.3390/s23073595.

[46] Christopher M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. New York, NY: Springer Science+Business Media, LLC, 2006. 758 pp.

[47] Fan Bo, Jia Li, Weibing Wang, and Kaiyue Zhou. "Robust Attitude and Heading Estimation under Dynamic Motion and Magnetic Disturbance". In: *Micromachines* 14.5 (May 2023), p. 1070. DOI: 10.3390/mi14051070.

[48] Gregory C. Bogdanis. "Effects of Physical Activity and Inactivity on Muscle Fatigue". In: *Frontiers in Physiology* 3 (2012). DOI: 10.3389/fphys.2012.00142.

[49] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. "Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models". In: (2021). DOI: 10.48550/ARXIV.2103.04922.

[50] Gunnar Borg. *Borg's Perceived exertion and pain scales*. Human Kinetics, 1998, p. 104.

[51] Gunnar Borg. "Psychophysical bases of perceived exertion". In: *Medicine and science in sports and exercise* 14.5 (1982), pp. 377–381.

[52] Yassine Bouteraa, Ismail Ben Abdallah, and Khalil Boukthir. "A New Wrist–Forearm Rehabilitation Protocol Integrating Human Biomechanics and SVM-Based Machine Learning for Muscle Fatigue Estimation". In: *Bioengineering* 10.2 (Feb. 2023), p. 219. DOI: 10.3390/bioengineering10020219.

[53] Louise Brennan, Antonio Bevilacqua, Tahar Kechadi, and Brian Caulfield. "Segmentation of shoulder rehabilitation exercises for single and multiple inertial sensor systems". In: *Journal of Rehabilitation and Assistive Technologies Engineering* 7 (Jan. 2020), p. 205566832091537. DOI: 10.1177/2055668320915377.

[54] 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*. IEEE, Aug. 2010. DOI: 10.1109/icpr.2010.764.

[55] C. Buckley, M.A. O'Reilly, D. Whelan, et al. "Binary classification of running fatigue using a single inertial measurement unit". In: *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, May 2017. DOI: 10.1109/bsn.2017.7936040.

[56] Achim Buerkle, Harveen Matharu, Ali Al-Yacoub, et al. "An adaptive human sensor framework for human–robot collaboration". In: *The International Journal of Advanced Manufacturing Technology* 119.1-2 (Nov. 2021), pp. 1233–1248. DOI: 10.1007/s00170-021-08299-2.

[57] Andreas Bulling, Ulf Blanke, and Bernt Schiele. "A tutorial on human activity recognition using body-worn inertial sensors". In: *ACM Computing Surveys* 46.3 (Jan. 2014), pp. 1–33. DOI: 10.1145/2499621.

[58] Denisse Bustos, Filipa Cardoso, Manoel Rios, et al. "Machine Learning Approach to Model Physical Fatigue during Incremental Exercise among Firefighters". In: *Sensors* 23.1 (Dec. 2022), p. 194. DOI: 10.3390/s23010194.

[59] Eglė Butkevičiūtė, Matīss Eriņš, and Liepa Bikulčienė. “An adaptable human fatigue evaluation system”. In: *Procedia Computer Science* 192 (2021), pp. 1274–1284. DOI: 10.1016/j.procs.2021.08.131.

[60] Kelsy Cabello-Solorzano, Isabela Ortigosa de Araujo, Marco Peña, Luís Correia, and Antonio J. Tallón-Ballesteros. “The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis”. In: *Lecture Notes in Networks and Systems*. Springer Nature Switzerland, 2023, pp. 344–353. DOI: 10.1007/978-3-031-42536-3_33.

[61] Thomas W. Calvert, Eric W. Banister, Margaret V. Savage, and Tim Bach. “A Systems Model of the Effects of Training on Physical Performance”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.2 (Feb. 1976), pp. 94–102. DOI: 10.1109/tsmc.1976.5409179.

[62] Fernando Camarena, Miguel Gonzalez-Mendoza, Leonardo Chang, and Ricardo Cuevas-Ascencio. “An Overview of the Vision-Based Human Action Recognition Field”. In: *Mathematical and Computational Applications* 28.2 (Apr. 2023), p. 61. DOI: 10.3390/mca28020061.

[63] Manuel Lage Cañellas, Constantino Álvarez Casado, Le Nguyen, and Miguel Bordallo López. *Estimating exercise-induced fatigue from thermal facial images*. 2023. DOI: 10.48550/ARXIV.2309.06095.

[64] W. B. Cannon. “The wisdom of the body”. In: *Oxford, England: Norton & Co.* (1939).

[65] D. L. Carey, K. Ong, M. E. Morris, J. Crow, and K. M. Crossley. “Predicting ratings of perceived exertion in Australian football players: methods for live estimation”. In: *International Journal of Computer Science in Sport* 15.2 (Dec. 2016), pp. 64–77. DOI: 10.1515/ijcss-2016-0005.

[66] Sang Hoon Chae, Yushin Kim, Kyoung-Soub Lee, and Hyung-Soon Park. “Development and Clinical Evaluation of a Web-Based Upper Limb Home Rehabilitation System Using a Smartwatch and Machine Learning Model for Chronic Stroke Survivors: Prospective Comparative Study”. In: *JMIR mHealth and uHealth* 8.7 (July 2020), e17216. DOI: 10.2196/17216.

[67] Christos Chalitsios, Thomas Nikodelis, Vasileios Konstantakos, and Iraklis Kollias. “Sensitivity of movement features to fatigue during an exhaustive treadmill run”. In: *European Journal of Sport Science* 22.9 (July 2021), pp. 1374–1382. DOI: 10.1080/17461391.2021.1955015.

[68] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”. In: (2011). DOI: 10.48550/ARXIV.1106.1813.

[69] Yeok Tatt Cheah, Ka Wing Frances Wan, and Joanne Yip. “Prediction of Muscle Fatigue During Dynamic Exercises based on Surface Electromyography Signals Using Gaussian Classifier”. In: *Physical Ergonomics and Human Factors*. AHFE International, 2022. DOI: 10.54941/ahfe1002597.

[70] Shoukun Chen, Kaili Xu, Xiwen Yao, et al. “Information fusion and multi-classifier system for miner fatigue recognition in plateau environments based on electrocardiography and electromyography signals”. In: *Computer Methods and Programs in Biomedicine* 211 (Nov. 2021), p. 106451. DOI: 10.1016/j.cmpb.2021.106451.

[71] Shoukun Chen, Kaili Xu, Xiwen Yao, et al. “Psychophysiological data-driven multi-feature information fusion and recognition of miner fatigue in high-altitude and cold areas”. In: *Computers in Biology and Medicine* 133 (June 2021), p. 104413. DOI: 10.1016/j.compbiomed.2021.104413.

[72] Xilai Chen, Meiqin Liu, and Senlin Zhang. “An LSTM-Attention-based Method to Muscle Fatigue Detection by Integrating Multi-Source sEMG Signals”. In: *2021 40th Chinese Control Conference (CCC)*. IEEE, July 2021. DOI: 10.23919/ccc52363.2021.9549359.

[73] Michael R. Chernick. *Bootstrap methods. A guide for practitioners and researchers*. 2nd ed. Wiley series in probability and statistics. Includes bibliographical references (p. 188-329) and indexes. Hoboken, NJ: Wiley-Interscience, 2008. 369 pp.

[74] Yiu-Ming Cheung and Hong Jia. “Unsupervised Feature Selection with Feature Clustering”. In: *WI-IAT ’12: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. ACM Conferences. Washington, DC: IEEE Computer Society, Dec. 2012, pp. 9–15.

[75] Davide Chicco, Luca Oneto, and Erica Tavazzi. “Eleven quick tips for data cleaning and feature engineering”. In: *PLOS Computational Biology* 18.12 (Dec. 2022). Ed. by Francis Ouellette, e1010718. DOI: 10.1371/journal.pcbi.1010718.

[76] Anirban Dutta Choudhury, Rohan Banerjee, Sanjay Kimbahune, and Arpan Pal. “Sensor signal analytics”. In: *New Frontiers of Cardiovascular Screening Using Unobtrusive Sensors, AI, and IoT*. Elsevier, 2022, pp. 61–89. DOI: 10.1016/b978-0-12-824499-9.00003-9.

[77] Alok Kumar Chowdhury, Dian Tjondronegoro, Vinod Chandran, Jinglan Zhang, and Stewart G. Trost. “Prediction of Relative Physical Activity Intensity Using Multimodal Sensing of Physiological Data”. In: *Sensors* 19.20 (Oct. 2019), p. 4509. DOI: 10.3390/s19204509.

[78] Dhiman Deb Chowdhury. “Fundamentals of Time Synchronization”. In: *NextGen Network Synchronization*. Springer International Publishing, 2021, pp. 1–16. DOI: 10.1007/978-3-030-71179-5_1.

[79] Seungeun Chung, Jiyoun Lim, Kyoung Ju Noh, Gague Kim, and Hyuntae Jeong. “Sensor Data Acquisition and Multimodal Sensor Fusion for Human Activity Recognition Using Deep Learning”. In: *Sensors* 19.7 (Apr. 2019), p. 1716. DOI: 10.3390/s19071716.

[80] Marinella Coco, Andrea Buscemi, Tiziana Ramaci, et al. “Influences of Blood Lactate Levels on Cognitive Domains and Physical Health during a Sports Stress. Brief Review”. In: *International Journal of Environmental Research and Public Health* 17.23 (Dec. 2020), p. 9043. DOI: 10.3390/ijerph17239043.

[81]Jussi Collin, Pavel Davidson, Martti Kirkko-Jaakkola, and Helena Leppäkoski. "Inertial Sensors and Their Applications". In: *Handbook of Signal Processing Systems*. Springer International Publishing, Oct. 2018, pp. 51–85. DOI: 10.1007/978-3-319-91734-4_2.

[82]Paul Comfort, John J. McMahon, and Timothy J. Suchomel. "Optimizing Squat Technique – Revisited". In: *Strength & Conditioning Journal* 40.6 (Dec. 2018), pp. 68–74. DOI: 10.1519/ssc.0000000000000398.

[83]Elsa Concha-Pérez, Hugo G. Gonzalez-Hernandez, and Jorge A. Reyes-Avendaño. "Physical Exertion Recognition Using Surface Electromyography and Inertial Measurements for Occupational Ergonomics". In: *Sensors* 23.22 (Nov. 2023), p. 9100. DOI: 10.3390/s23229100.

[84]Jonathan Cook and Vikram Ramadas. "When to consult precision-recall curves". In: *The Stata Journal: Promoting communications on statistics and Stata* 20.1 (Mar. 2020), pp. 131–148. DOI: 10.1177/1536867x20909693.

[85]Giorgio Corani and Marco Zaffalon. "Learning Reliable Classifiers From Small or Incomplete Data Sets: The Naive Credal Classifier 2". In: *The Journal of Machine Learning Research* 9 (June 2008), pp. 581–621.

[86]Juliet M. Corbin. *Basics of qualitative research. Techniques and procedures for developing grounded theory*. Ed. by Anselm L. Strauss. Fourth edition. Literaturverzeichnis: Seiten 413-419. Los Angeles: SAGE, 2015. 431 pp.

[87]Edward F Coyle. "Physical activity as a metabolic stressor". In: *The American Journal of Clinical Nutrition* 72.2 (Aug. 2000), 512S–520S. DOI: 10.1093/ajcn/72.2.512s.

[88]John W. Creswell. *Research design. Qualitative, quantitative, and mixed methods approaches*. Ed. by J. David Creswell. Sixth edition, international student edition. Literaturverzeichnis: Seiten 271-278. Los Angeles: Sage, 2023. 291 pp.

[89]Padraig Cunningham, Bahavathy Kathirgamanathan, and Sarah Jane Delany. "Feature Selection Tutorial with Python Examples". In: (2021). DOI: 10.48550/ARXIV.2106.06437.

[90]Victoria Da Poian, Bethany Theiling, Lily Clough, et al. "Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry". In: *Frontiers in Astronomy and Space Sciences* 10 (May 2023). DOI: 10.3389/fspas.2023.1134141.

[91]L. Minh Dang, Kyungbok Min, Hanxiang Wang, et al. "Sensor-based and vision-based human activity recognition: A comprehensive survey". In: *Pattern Recognition* 108 (Dec. 2020), p. 107561. DOI: 10.1016/j.patcog.2020.107561.

[92]Yukun Dang, Zitong Liu, Xixin Yang, Linqiang Ge, and Sheng Miao. "A fatigue assessment method based on attention mechanism and surface electromyography". In: *Internet of Things and Cyber-Physical Systems* 3 (2023), pp. 112–120. DOI: 10.1016/j.iotcps.2023.03.002.

[93] A.M. Danilova, E.V. Filushkina, and A.D. Voronin. “Application of Digital Technologies During Remote Training Lessons”. In: *Izvestiya of the Samara Science Centre of the Russian Academy of Sciences. Social, Humanitarian, Medicobiologial Sciences* 83 (2022), pp. 15–19. DOI: 10.37313/2413-9645-2022-24-83-15-19.

[94] Hamed Darbandi, Carolien Munsters, and Paul Havinga. “Non-Invasive Lactate Estimation Using Wearable Sensors for Remote Fatigue Assessment in Horses”. In: *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, Mar. 2024. DOI: 10.1109/percomworkshops59983.2024.10503247.

[95] Padraig Davidson, Peter Düking, Christoph Zinner, Billy Sperlich, and Andreas Hotho. “Smartwatch-Derived Data and Machine Learning Algorithms Estimate Classes of Ratings of Perceived Exertion in Runners: A Pilot Study”. In: *Sensors* 20.9 (May 2020), p. 2637. DOI: 10.3390/s20092637.

[96] Valentina De Simone, Valentina Di Pasquale, and Salvatore Miranda. “An overview on the use of AI/ML in Manufacturing MSMEs: solved issues, limits, and challenges”. In: *Procedia Computer Science* 217 (2023), pp. 1820–1829. DOI: 10.1016/j.procs.2022.12.382.

[97] Luca De Vito, Enrico Picariello, Francesco Picariello, et al. “Measurement System for Operator 5.0: a Learning Fatigue Recognition based on sEMG Signals”. In: *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. Vol. pp. IEEE, June 2023, pp. 1–6. DOI: 10.1109/memea57477.2023.10171933.

[98] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. “Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey”. In: *IEEE Access* 8 (2020), pp. 210816–210836. DOI: 10.1109/access.2020.3037715.

[99] Miral Desai and Hiren Mewada. “A novel approach for yoga pose estimation based on in-depth analysis of human body joint detection accuracy”. In: *PeerJ Computer Science* 9 (Jan. 2023), e1152. DOI: 10.7717/peerj-cs.1152.

[100] Aayush Dhattarwal and Saroj Ratnoo. “A Review of Deep Learning Techniques for Human Activity Recognition”. In: *Lecture Notes in Networks and Systems*. Springer Nature Switzerland, 2023, pp. 313–327. DOI: 10.1007/978-3-031-27409-1_28.

[101] Genny Dimitrakopoulou, Nikolaos Kapsalis, and George Kokkinis. “Bias Detection and Correction Methods for Machine Learning Algorithms”. In: *2024 5th International Conference in Electronic Engineering, Information Technology & Education (EEITE)*. Vol. 2. IEEE, May 2024, pp. 1–6. DOI: 10.1109/eeite61750.2024.10654404.

[102] Hannah L. Dimmick, Cody R. van Rassel, Martin J. MacInnis, and Reed Ferber. “Use of subject-specific models to detect fatigue-related changes in running biomechanics: a random forest approach”. In: *Frontiers in Sports and Active Living* 5 (Dec. 2023). DOI: 10.3389/fspor.2023.1283316.

[103]Giovanni Diraco, Gabriele Rescio, Pietro Siciliano, and Alessandro Leone. “Review on Human Action Recognition in Smart Living: Sensing Technology, Multimodality, Real-Time Processing, Interoperability, and Resource-Constrained Processing”. In: *Sensors* 23.11 (June 2023), p. 5281. DOI: 10.3390/s23115281.

[104]Shengshun Duan, Qiongfeng Shi, and Jun Wu. “Multimodal Sensors and ML-Based Data Fusion for Advanced Robots”. In: *Advanced Intelligent Systems* 4.12 (Oct. 2022). DOI: 10.1002/aisy.202200213.

[105]Ralf Eggeling, Mikko Koivisto, and Ivo Grosse. “Dealing with small data: on the generalization of context trees”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning* 37 (July 2015), pp. 1245–1253.

[106]George Ellis. “Filters in Control Systems”. In: *Control System Design Guide*. Elsevier, 2012, pp. 165–183. DOI: 10.1016/b978-0-12-385920-4.00009-6.

[107]Ahmed Elsais, Vegard B Wyller, Jon Håvard Loge, and Emilia Kerty. “Fatigue in myasthenia gravis: is it more than muscular weakness?” In: *BMC Neurology* 13.1 (Oct. 2013). DOI: 10.1186/1471-2377-13-132.

[108]Mohamed Elshafei, Diego Elias Costa, and Emad Shihab. “On the Impact of Biceps Muscle Fatigue in Human Activity Recognition”. In: *Sensors* 21.4 (2021). DOI: 10.3390/s21041070.

[109]Roger M. Enoka and Jacques Duchateau. “Muscle fatigue: what, why and how it influences muscle function”. In: *The Journal of Physiology* 586.1 (Jan. 2008), pp. 11–23. DOI: 10.1113/jphysiol.2007.139477.

[110]Roger M. Enoka and Jacques Duchateau. “Translating Fatigue to Human Performance”. In: *Medicine & Science in Sports & Exercise* 48.11 (Nov. 2016), pp. 2228–2238. DOI: 10.1249/mss.0000000000000929.

[111]Elissa S. Epel, Alexandra D. Crosswell, Stefanie E. Mayer, et al. “More than a feeling: A unified view of stress measurement for population science”. In: *Frontiers in Neuroendocrinology* 49 (Apr. 2018), pp. 146–169. DOI: 10.1016/j.yfrne.2018.03.001.

[112]Elena Escobar-Linero, Manuel Domínguez-Morales, and José Luis Sevillano. “Worker’s physical fatigue classification using neural networks”. In: *Expert Systems with Applications* 198 (July 2022), p. 116784. DOI: 10.1016/j.eswa.2022.116784.

[113]Roger Eston. “Use of Ratings of Perceived Exertion in Sports”. In: *International Journal of Sports Physiology and Performance* 7.2 (June 2012), pp. 175–182. DOI: 10.1123/ijspp.7.2.175.

[114]Daniel R. Evans, Ian A. Boggero, and Suzanne C. Segerstrom. “The Nature of Self-Regulatory Fatigue and Ego Depletion”. In: *Personality and Social Psychology Review* 20.4 (June 2016), pp. 291–310. DOI: 10.1177/1088868315597841.

[115]Julian J. Faraway and Nicole H. Augustin. “When small data beats big data”. In: *Statistics & Probability Letters* 136 (May 2018), pp. 142–145. DOI: 10.1016/j.spl.2018.02.031.

[116] Weibin Feng, Kelong Zeng, Xiaomei Zeng, et al. “Predicting physical fatigue in athletes in rope skipping training using ECG signals”. In: *Biomedical Signal Processing and Control* 83 (May 2023), p. 104663. DOI: 10.1016/j.bspc.2023.104663.

[117] Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V. Chawla. “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary”. In: *Journal Artificial Interna Res.* 61.1 (Jan. 2018), pp. 863–905.

[118] Dalson Britto Figueiredo Filho, José Alexandre Silva Júnior, and Enivaldo Carvalho Rocha. “What is R2 all about?” In: *Leviathan (São Paulo)* 3 (Nov. 2011), p. 60. DOI: 10.11606/issn.2237-4485.lev.2011.132282.

[119] George Fink. *Stress. Handbook of Stress Series, Volume 1*. San Diego, 2016.

[120] Michael Fire and Carlos Guestrin. “Over-optimization of academic publishing metrics: observing Goodhart’s Law in action”. In: *GigaScience* 8.6 (May 2019). DOI: 10.1093/gigascience/giz053.

[121] Jeannie Fitzgerald and Conor Ryan. “On size, complexity and generalisation error in GP”. In: *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. GECCO ’14. ACM, July 2014. DOI: 10.1145/2576768.2598346.

[122] George Forman and Ira Cohen. “Learning from Little: Comparison of Classifiers Given Little Training”. In: *Knowledge Discovery in Databases: PKDD 2004*. Springer Berlin Heidelberg, 2004, pp. 161–172. DOI: 10.1007/978-3-540-30116-5_17.

[123] John Frank, Rosemary Foster, and Claudia Pagliari. “Open access publishing – noble intention, flawed reality”. In: *Social Science & Medicine* 317 (Jan. 2023), p. 115592. DOI: 10.1016/j.socscimed.2022.115592.

[124] Biying Fu, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. “Sensing Technology for Human Activity Recognition: A Comprehensive Survey”. In: *IEEE Access* 8 (2020), pp. 83791–83820. DOI: 10.1109/access.2020.2991891.

[125] Reinhard Fuchs and Markus Gerber, eds. *Handbuch Stressregulation und Sport*. Springer Berlin Heidelberg, 2018. DOI: 10.1007/978-3-662-49322-9.

[126] Luoyu Gan, Zhaoyang Yang, Yanfei Shen, et al. “Heart rate variability analysis method for exercise-induced fatigue monitoring”. In: *Biomedical Signal Processing and Control* 92 (June 2024), p. 105966. DOI: 10.1016/j.bspc.2024.105966.

[127] Guillermo García Pérez de Sevilla, Olga Barceló Guido, María de la Paz De la Cruz, et al. “Remotely Supervised Exercise during the COVID-19 Pandemic versus in-Person-Supervised Exercise in Achieving Long-Term Adherence to a Healthy Lifestyle”. In: *International Journal of Environmental Research and Public Health* 18.22 (Nov. 2021), p. 12198. DOI: 10.3390/ijerph182212198.

[128] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. Concepts, tools, and techniques to build intelligent systems*. Third edition. Die 1. edition erschien 2017 unter dem Titel "Hands-on machine learning with Scikit-Learn and TensorFlow". Beijing: O’Reilly, 2023. 834 pp.

[129]Youri Geurkink, Gilles Vandewiele, Maarten Lievens, et al. “Modeling the Prediction of the Session Rating of Perceived Exertion in Soccer: Unraveling the Puzzle of Predictive Indicators”. In: *International Journal of Sports Physiology and Performance* 14.6 (July 2019), pp. 841–846. DOI: 10.1123/ijspp.2018-0698.

[130]Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, et al. “Review on Psychological Stress Detection Using Biosignals”. In: *IEEE Transactions on Affective Computing* 13.1 (Jan. 2019), pp. 440–460. DOI: 10.1109/taffc.2019.2927337.

[131]Oonagh M Giggins, Kevin T Sweeney, and Brian Caulfield. “Rehabilitation exercise assessment using inertial sensors: a cross-sectional analytical study”. In: *Journal of NeuroEngineering and Rehabilitation* 11.1 (2014), p. 158. DOI: 10.1186/1743-0003-11-158.

[132]Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. “Monitoring stress with a wrist device using context”. In: *Journal of Biomedical Informatics* 73 (Sept. 2017), pp. 159–170. DOI: 10.1016/j.jbi.2017.08.006.

[133]Lívea Dornela Godoy, Matheus Teixeira Rossignoli, Polianna Delfino-Pereira, Norberto Garcia-Cairasco, and Eduardo Henrique de Lima Umeoka. “A Comprehensive Overview on Stress Neurobiology: Basic Concepts and Clinical Implications”. In: *Frontiers in Behavioral Neuroscience* 12 (July 2018). DOI: 10.3389/fnbeh.2018.00127.

[134]Markus Goldstein and Seiichi Uchida. “A Comparative Study on Outlier Removal from a Large-scale Dataset using Unsupervised Anomaly Detection”. In: *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science, 2016. DOI: 10.5220/0005701302630269.

[135]Yuri Gordienko, Sergii Stirenko, Yuriy Kochura, et al. *Deep Learning for Fatigue Estimation on the Basis of Multimodal Human-Machine Interactions*. 2018. DOI: 10.48550/ARXIV.1801.06048.

[136]Aishwarya Goyal, Shailendra Singh, Dharam Vir, and Dwarka Pershad. “Automation of Stress Recognition Using Subjective or Objective Measures”. In: *Psychological Studies* 61.4 (Nov. 2016), pp. 348–364. DOI: 10.1007/s12646-016-0379-1.

[137]M. Guaitolini, L. Truppa, A. M. Sabatini, A. Mannini, and C. Castagna. “Sport-induced fatigue detection in gait parameters using inertial sensors and support vector machines”. In: *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, Nov. 2020. DOI: 10.1109/biorob49111.2020.9224449.

[138]Xiaole Guan, Yanfei Lin, Qun Wang, Zhiwen Liu, and Chengyi Liu. “Sports fatigue detection based on deep learning”. In: *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, Oct. 2021. DOI: 10.1109/cisp-bmei53629.2021.9624395.

[139]Hao Guo, Li Ke, Qiang Du, and Song Guo. “Muscle fatigue state classification based on blood flow bioimpedance”. In: *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, Nov. 2022. DOI: 10.1109/cisp-bmei56279.2022.9980152.

[140]Xiaonan Guo, Jian Liu, and Yingying Chen. “When your wearables become your fitness mate”. In: *Smart Health* 16 (May 2020), p. 100114. DOI: 10.1016/j.smhl.2020.100114.

[141]Anthony C Hackney. “Stress and the neuroendocrine system: the role of exercise as a stressor and modifier of stress”. In: *Expert Review of Endocrinolog & Metabolism* 1.6 (Nov. 2006), pp. 783–792. DOI: 10.1586/17446651.1.6.783.

[142]Anthony C. Hackney and Amy R. Lane. “Exercise and the Regulation of Endocrine Hormones”. In: *Progress in Molecular Biology and Translational Science*. Elsevier, 2015, pp. 293–311. DOI: 10.1016/bs.pmbts.2015.07.001.

[143]Shona L. Halson. “Monitoring Training Load to Understand Fatigue in Athletes”. In: *Sports Medicine* 44.S2 (Sept. 2014), pp. 139–147. DOI: 10.1007/s40279-014-0253-z.

[144]Michael John Hamlin, Danielle Wilkes, Catherine A. Elliot, Catherine A. Lizamore, and Yaso Kathiravel. “Monitoring Training Loads and Perceived Stress in Young Elite University Athletes”. In: *Frontiers in Physiology* 10 (Jan. 2019). DOI: 10.3389/fphys.2019.00034.

[145]Steve Hanneke. “The Optimal Sample Complexity of PAC Learning”. In: (2015). DOI: 10.48550/ARXIV.1507.00473.

[146]Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. *Liquid Time-constant Networks*. 2020. DOI: 10.48550/arxiv.2006.04439.

[147]Mohammad Mehedi Hassan, Shamsul Huda, Md Zia Uddin, Ahmad Almogren, and Majed Alrubaian. “Human Activity Recognition from Body Sensor Data using Deep Learning”. In: *Journal of Medical Systems* 42.6 (Apr. 2018). DOI: 10.1007/s10916-018-0948-z.

[148]Hossein Hassani, Christina Beneki, Emmanuel Sirimal Silva, Nicolas Vandeput, and Dag Øivind Madsen. “The science of statistics versus data science: What is the future?” In: *Technological Forecasting and Social Change* 173 (Dec. 2021), p. 121111. DOI: 10.1016/j.techfore.2021.121111.

[149]Trevor Hastie. *The elements of statistical learning. Data mining, inference, and prediction*. Ed. by Robert Tibshirani and Jerome H. Friedman. Second edition. Springer Series in Statistics. Description based on publisher supplied metadata and other sources. New York, NY: Springer, 2009. 1745 pp.

[150]Eric B. Hekler, Predrag Klasnja, Guillaume Chevance, et al. “Why we need a small data paradigm”. In: *BMC Medicine* 17.1 (July 2019). DOI: 10.1186/s12916-019-1366-x.

[151]D.H. Hellhammer, A.A. Stone, J. Hellhammer, and J. Broderick. “Measuring Stress”. In: *Encyclopedia of Behavioral Neuroscience*. Elsevier, 2010, pp. 186–191. DOI: 10.1016/b978-0-08-045396-5.00188-3.

[152]A. Ralph Henderson. “The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data”. In: *Clinica Chimica Acta* 359.1–2 (Sept. 2005), pp. 1–26. DOI: 10.1016/j.cccn.2005.04.002.

[153]Robin Hermann, Daniel Lay, Patrick Wahl, Walton T. Roth, and Katja Petrowski. "Effects of psychosocial and physical stress on lactate and anxiety levels". In: *Stress* 22.6 (May 2019), pp. 664–669. DOI: 10.1080/10253890.2019.1610743.

[154]Andrew Hoblitzell, Meghna Babbar-Sebens, and Snehasis Mukhopadhyay. "Machine Learning with Small Data for User Modeling of Watershed Stakeholders Engaged in Interactive Optimization". In: *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*. CSAI '18. ACM, Dec. 2018. DOI: 10.1145/3297156.3297207.

[155]Niklas Hoefflin, Tim Spula, André Jeworutzki, and Jan Schwarzer. "Real-time Lateral Sitting Posture Detection using YOLOv5". In: *IEEE RAS EMBS 10th International Conference on Biomedical Robotics and Biomechatronics (BioRob 2024)* (Sept. 2024), p. 5. DOI: 10.1109/biorob60516.2024.10719953.

[156]Jin-Hyuk Hong, Julian Ramos, and Anind K. Dey. "Understanding physiological responses to stressors during physical activity". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, Sept. 2012. DOI: 10.1145/2370216.2370260.

[157]Rohit Hooda, Vedant Joshi, and Manan Shah. "A comprehensive review of approaches to detect fatigue using machine learning techniques". In: *Chronic Diseases and Translational Medicine* 8.1 (Feb. 2022), pp. 26–35. DOI: 10.1016/j.cdtm.2021.07.002.

[158]Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey". In: *ACM Journal on Responsible Computing* 1.2 (June 2024), pp. 1–52. DOI: 10.1145/3631326.

[159]Karen Hovsepian, Mustafa al'Absi, Emre Ertin, et al. "cStress". In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Sept. 2015. DOI: 10.1145/2750858.2807526.

[160]Lei Huang, Jie Qin, Yi Zhou, et al. "Normalization Techniques in Training DNNs: Methodology, Analysis and Application". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.8 (Aug. 2023), pp. 10173–10196. DOI: 10.1109/tpami.2023.3250241.

[161]Yanyan Huang, Chenyang Wang, Yifeng Sun, and Zhikun Jin. "Classifying static muscle fatigue: an surface electromyography signal analysis approach". In: *Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024)*. Ed. by Zulqarnain Baloch and Pier Paolo Piccaluga. SPIE, July 2024. DOI: 10.1117/12.3036611.

[162]Zawar Hussain, Quan Z. Sheng, and Wei Emma Zhang. "A review and categorization of techniques on device-free human activity recognition". In: *Journal of Network and Computer Applications* 167 (Oct. 2020), p. 102738. DOI: 10.1016/j.jnca.2020.102738.

[163]Frank Imbach, Nicolas Sutton-Charani, Jacky Montmain, Robin Candau, and Stéphane Perrey. “The Use of Fitness-Fatigue Models for Sport Performance Modelling: Conceptual Issues and Contributions from Machine-Learning”. In: *Sports Medicine - Open* 8.1 (Mar. 2022). DOI: 10.1186/s40798-022-00426-x.

[164]Franco M. Impellizzeri, Samuele M. Marcora, and Aaron J. Coutts. “Internal and External Training Load: 15 Years On”. In: *International Journal of Sports Physiology and Performance* 14.2 (Feb. 2019), pp. 270–273. DOI: 10.1123/ijspp.2018-0935.

[165]Hanifatul Insan, Sri Suryani Prasetiyowati, and Yuliant Sibaroni. “SMOTE-LOF and Borderline-SMOTE Performance to Overcome Imbalanced Data and Outliers on Classification”. In: *2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*. IEEE, Dec. 2023. DOI: 10.1109/icicyta60173.2023.10428902.

[166]Brian Kenji Iwana and Seiichi Uchida. “An empirical survey of data augmentation for time series classification with neural networks”. In: *PLOS ONE* 16.7 (July 2021). Ed. by Friedhelm Schwenker, e0254841. DOI: 10.1371/journal.pone.0254841.

[167]Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, and Thomas S. Huang. “Guest Editors’ Introduction: Human-Centered Computing–Toward a Human Revolution”. In: *Computer* 40.5 (May 2007), pp. 30–34. DOI: 10.1109/mc.2007.169.

[168]Ashish Jaiswal, Mohammad Zaki Zadeh, Aref Hebri, and Fillia Makedon. *Assessing Fatigue with Multimodal Wearable Sensors and Machine Learning*. 2022. DOI: 10.48550/ARXIV.2205.00287.

[169]Fauzani.N Jamaluddin, Siti A. Ahmad, Samsul Bahari Mohd Noor, Wan Zuha Wan Hassan, and E.F Shair. “Performance of Different Threshold Estimation Methods on SEMG Wavelet De-noising in Prolonged Fatigue Identification”. In: *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, Dec. 2018. DOI: 10.1109/iecbes.2018.8626599.

[170]Daniel Janssen, Wolfgang I. Schöllhorn, Karl M. Newell, et al. “Diagnosing fatigue in gait patterns by support vector machines and self-organizing maps”. In: *Human Movement Science* 30.5 (Oct. 2011), pp. 966–975. DOI: 10.1016/j.humov.2010.08.010.

[171]Houtan Jebelli, Byungjoo Choi, and SangHyun Lee. “Application of Wearable Biosensors to Construction Sites. II: Assessing Workers’ Physical Demand”. In: *Journal of Construction Engineering and Management* 145.12 (Dec. 2019). DOI: 10.1061/(asce)co.1943-7862.0001710.

[172]André Jeworutzki. “Determining acute physiological stress levels with wearable sensors based on movement quality and exhaustion during repetitive training exercises”. In: *Designing Interactive Systems Conference*. DIS ’22. ACM, June 2022. DOI: 10.1145/3532107.3532874.

[173] Andre Jeworutzki, Jan Schwarzer, Kai von Luck, Susanne Draheim, and Qi Wang. "A Preliminary Experimental Outline to Train Machine Learning Models for the Unobtrusive, Real-Time Detection of Acute Physiological Stress Levels during Training Exercises". In: *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*. PETRA '21. ACM, June 2021. DOI: 10.1145/3453892.3461833.

[174] André Jeworutzki, Jan Schwarzer, Kai Von Luck, et al. "Small Data, Big Challenges: Pitfalls and Strategies for Machine Learning in Fatigue Detection". In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, June 2023. DOI: 10.1145/3594806.3594825.

[175] Yanran Jiang, Vincent Hernandez, Gentiane Venture, Dana Kulić, and Bernard K. Chen. "A Data-Driven Approach to Predict Fatigue in Exercise Based on Motion Data from Wearable Sensors or Force Plate". In: *Sensors* 21.4 (Feb. 2021), p. 1499. DOI: 10.3390/s21041499.

[176] Yanran Jiang, Peter Malliaras, Bernard Chen, and Dana Kulić. "Real-time forecasting of exercise-induced fatigue from wearable sensors". In: *Computers in Biology and Medicine* 148 (Sept. 2022), p. 105905. DOI: 10.1016/j.compbiomed.2022.105905.

[177] Arlene John, Koushik Kumar Nundy, Barry Cardiff, and Deepu John. "Multimodal Multiresolution Data Fusion Using Convolutional Neural Networks for IoT Wearable Sensing". In: *IEEE Transactions on Biomedical Circuits and Systems* 15.6 (Dec. 2021), pp. 1161–1173. DOI: 10.1109/tbcas.2021.3134043.

[178] Jack W Judy. "Microelectromechanical systems (MEMS): fabrication, design and applications". In: *Smart Materials and Structures* 10.6 (Nov. 2001), pp. 1115–1134. DOI: 10.1088/0964-1726/10/6/301.

[179] Alexander Jung. *Machine learning. The basics*. Machine learning: foundations, methodologies, and applications. Singapore: Springer Singapore, 2022. 212 pp.

[180] Patrick Juola. "Data Integrity". In: *Encyclopedia of Big Data*. Springer International Publishing, 2022, pp. 294–295. DOI: 10.1007/978-3-319-32010-6_307.

[181] Divya Bharathi K, Karthick P. A., and Ramakrishnan S. "Automated detection of muscle fatigue conditions from cyclostationary based geometric features of surface electromyography signals". In: *Computer Methods in Biomechanics and Biomedical Engineering* 25.3 (July 2021), pp. 320–332. DOI: 10.1080/10255842.2021.1955104.

[182] Daniel Kahneman. *Noise. A flaw in human judgment*. Ed. by Olivier Sibony and Cass R. Sunstein. First Little, Brown Spark paperback edition. New York, NY: Little, Brown Spark, 2022. 452 pp.

[183] Manasa Kalanadhabhatta, Chulhong Min, Alessandro Montanari, and Fahim Kawsar. "FatigueSet: A Multi-modal Dataset for Modeling Mental Fatigue and Fatigability". In: *Pervasive Computing Technologies for Healthcare*. Springer International Publishing, 2022, pp. 204–217. DOI: 10.1007/978-3-030-99194-4_14.

[184]Johanna Kallio, Elena Vildjiounaite, Jani Koivusaari, et al. “Assessment of perceived indoor environmental quality, stress and productivity based on environmental sensor data and personality categorization”. In: *Building and Environment* 175 (May 2020), p. 106787. DOI: 10.1016/j.buildenv.2020.106787.

[185]Michelle Karg, Kolja Kühnlenz, Martin Buss, et al. “Expression and Automatic Recognition of Exhaustion in Natural Walking”. In: *Interfaces and Human Computer Interaction (IHCI)* (Jan. 2008).

[186]Michelle Karg, Gentiane Venture, Jesse Hoey, and Dana Kulic. “Human Movement Analysis as a Measure for Fatigue: A Hidden Markov-Based Approach”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22.3 (May 2014), pp. 470–481. DOI: 10.1109/tnsre.2013.2291327.

[187]P.A. Karthick, Diptasree Maitra Ghosh, and S. Ramakrishnan. “Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms”. In: *Computer Methods and Programs in Biomedicine* 154 (Feb. 2018), pp. 45–56. DOI: 10.1016/j.cmpb.2017.10.024.

[188]Swapnali Karvekar, Masoud Abdollahi, and Ehsan Rashedi. “A Data-Driven Model to Identify Fatigue Level Based on the Motion Data from a Smartphone”. In: (Oct. 2019). DOI: 10.1101/796854.

[189]Swapnali Karvekar, Masoud Abdollahi, and Ehsan Rashedi. “Smartphone-based human fatigue level detection using machine learning approaches”. In: *Ergonomics* 64.5 (Jan. 2021), pp. 600–612. DOI: 10.1080/00140139.2020.1858185.

[190]Bahavathy Kathirgamanathan, Brian Caulfield, and Pádraig Cunningham. *Multivariate Time Series data of Fatigued and Non-Fatigued Running from Inertial Measurement Units*. en. 2023. DOI: 10.5281/ZENODO.7997850.

[191]Bahavathy Kathirgamanathan, Brian Caulfield, and Pádraig Cunningham. “Towards Globalised Models for Exercise Classification using Inertial Measurement Units”. In: *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*. Vol. 21. IEEE, Oct. 2023, pp. 1–4. DOI: 10.1109/bsn58485.2023.10331612.

[192]Bahavathy Kathirgamanathan and Pádraig Cunningham. “A Feature Selection Method for Multi-dimension Time-Series Data”. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2020, pp. 220–231. DOI: 10.1007/978-3-030-65742-0_15.

[193]S. Katz, T. D. Downs, H. R. Cash, and R. C. Grotz. “Progress in Development of the Index of ADL”. In: *The Gerontologist* 10 (Mar. 1970), pp. 20–30. DOI: 10.1093/geront/10.1_part_1.20.

[194]L. N. Keerthana and Abirami B. Sakthi. “Developing an Automatic Evaluation of Exertion Using a Smart Phone”. In: *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, Mar. 2023. DOI: 10.1109/icaccs57279.2023.10112834.

[195]R. Khokale, A. R. Panat, and Y. H. Gulhane. “Analysis of affective speech for fatigue detection”. In: *Proceedings of the International Conference and Workshop on Emerging Trends in Technology*. ACM, Feb. 2010. DOI: 10.1145/1741906.1741960.

[196]Vladimir Khorikov. *Unit Testing Principles, Practices, and Patterns*. Description based on publisher supplied metadata and other sources. New York: Manning Publications Co. LLC, 2020. 1252 pp.

[197]Seungjun Kim, Dan Tasse, and Anind K. Dey. "Making Machine-Learning Applications for Time-Series Sensor Data Graphical and Interactive". In: *ACM Transactions on Interactive Intelligent Systems* 7.2 (June 2017), pp. 1–30. DOI: 10.1145/2983924.

[198]Rob Kitchin and Tracey P. Lauriault. "Small data in the era of big data". In: *GeoJournal* 80.4 (Oct. 2014), pp. 463–475. DOI: 10.1007/s10708-014-9601-7.

[199]B. M. Kluger, L. B. Krupp, and R. M. Enoka. "Fatigue and fatigability in neurologic illnesses: Proposal for a unified taxonomy". In: *Neurology* 80.4 (Jan. 2013), pp. 409–416. DOI: 10.1212/wnl.0b013e31827f07be.

[200]Rüya D Kocalevent, Andreas Hinz, Elmar Brähler, and Burghard F Klapp. "Determinants of fatigue and stress". In: *BMC Research Notes* 4.1 (July 2011). DOI: 10.1186/1756-0500-4-238.

[201]C.E. Koch, B. Leinweber, B.C. Drengberg, C. Blaum, and H. Oster. "Interaction between circadian rhythms and stress". In: *Neurobiology of Stress* 6 (Feb. 2017), pp. 57–67. DOI: 10.1016/j.ynstr.2016.09.001.

[202]Peter Kokol, Marko Kokol, and Sašo Zagoranski. "Machine learning on small size samples: A synthetic knowledge synthesis". In: *Science Progress* 105.1 (Jan. 2022), p. 003685042110297. DOI: 10.1177/00368504211029777.

[203]Johnson Kolluri, Vinay Kumar Kotte, M.S.B. Phridviraj, and Shaik Razia. "Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method". In: *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, June 2020. DOI: 10.1109/icoei48184.2020.9142992.

[204]Shuangshuang Kong, Hui Wang, and Kaijun Wang. "Conservative Generalisation for Small Data Analytics –An Extended Lattice Machine Approach". In: *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, Dec. 2020. DOI: 10.1109/icmlc51923.2020.9469579.

[205]Vladislav Kovalevsky, Radhakrishnan Delhibabu, and Nataly Zhukova. "AutoML Framework for Physical Activities Recognition". In: *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*. IEEE, May 2024. DOI: 10.1109/aiiot58432.2024.10574646.

[206]Yousef Kowsar, Masud Moshtaghi, Eduardo Velloso, Lars Kulik, and Christopher Leckie. "Detecting unseen anomalies in weight training exercises". In: *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI '16*. ACM Press, 2016. DOI: 10.1145/3010915.3010941.

[207]Max Kuhn. *Applied predictive modeling*. Ed. by Kjell Johnson. Corrected at 5th printing. Description based upon print version of record. New York: Springer, 2016. 1600 pp.

[208] Max Kuhn. *Feature engineering and selection. A practical approach for predictive models*. Ed. by Kjell Johnson. Chapman & Hall / CRC data science series. Includes bibliographical references and index. Boca Raton: CRC Press, Taylor & Francis Group, 2020. 297 pp.

[209] Sakshi Kulkarni, Shubham Deshmukh, Favin Fernandes, Aniket Patil, and Vaishali Jabade. "PoseAnalyser: A Survey on Human Pose Estimation". In: *SN Computer Science* 4.2 (Jan. 2023). DOI: 10.1007/s42979-022-01567-2.

[210] Teerath Kumar, Alessandra Mileo, Rob Brennan, and Malika Bendechache. *Image Data Augmentation Approaches: A Comprehensive Survey and Future directions*. 2023. DOI: 10.48550/ARXIV.2301.02830.

[211] Stefan Kupschick, Marie Pendzich, Dorata Gardas, et al. "Predicting firefighters' exertion based on machine learning techniques". In: (2016). DOI: 10.21934/BAUA:FOCUS20161107.

[212] Jan Kuschan and Jörg Krüger. "Fatigue recognition in overhead assembly based on a soft robotic exosuit for worker assistance". In: *CIRP Annals* 70.1 (2021), pp. 9–12. DOI: 10.1016/j.cirp.2021.04.034.

[213] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, et al. "Detecting Moments of Stress from Measurements of Wearable Physiological Sensors". In: *Sensors* 19.17 (Sept. 2019), p. 3805. DOI: 10.3390/s19173805.

[214] Marc Lafon and Alexandre Thomas. *Understanding the Double Descent Phenomenon in Deep Learning*. 2024. DOI: 10.48550/ARXIV.2403.10459.

[215] K L Lamb, R G Eston, and D Corns. "Reliability of ratings of perceived exertion during progressive treadmill exercise." In: *British Journal of Sports Medicine* 33.5 (Oct. 1999), pp. 336–339. DOI: 10.1136/bjsm.33.5.336.

[216] Arsalan Lambay, Ying Liu, Phillip Morgan, and Ze Ji. "A Data-Driven Fatigue Prediction using Recurrent Neural Networks". In: *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, June 2021. DOI: 10.1109/hora52670.2021.9461377.

[217] Arsalan Lambay, Ying Liu, Phillip L. Morgan, and Ze Ji. "Machine learning assisted human fatigue detection, monitoring, and recovery: A Review". In: *Digital Engineering* 1 (June 2024), p. 100004. DOI: 10.1016/j.dte.2024.100004.

[218] Guillaume Lame. "Systematic Literature Reviews: An Introduction". In: *Proceedings of the Design Society: International Conference on Engineering Design* 1.1 (July 2019), pp. 1633–1642. DOI: 10.1017/dsi.2019.169.

[219] Leslie Lamport. "Time, clocks, and the ordering of events in a distributed system". In: *Communications of the ACM* 21.7 (July 1978), pp. 558–565. DOI: 10.1145/359545.359563.

[220] Gongjin Lan, Yu Wu, Fei Hu, and Qi Hao. "Vision-Based Human Pose Estimation via Deep Learning: A Survey". In: *IEEE Transactions on Human-Machine Systems* 53.1 (Feb. 2023), pp. 253–268. DOI: 10.1109/thms.2022.3219242.

[221]Bart De Langhe and Philip Fernbach. “The dangers of categorical thinking”. In: *Harvard Business Review 97* 5 (Sept. 2019), pp. 80–92.

[222]Tracey Lauriault. “Data, infrastructures and geographical imaginations”. PhD thesis. Ottawa: Carleton University, May 2012. DOI: 10.22215/etd/2012-09890.

[223]Richard S Lazarus and Susan Folkman. “Stress, appraisal, and coping”. In: *Springer publishing company* (1984).

[224]Martin Levesque and David Tipper. “A Survey of Clock Synchronization Over Packet-Switched Networks”. In: *IEEE Communications Surveys & Tutorials* 18.4 (2016), pp. 2926–2947. DOI: 10.1109/comst.2016.2590438.

[225]Dujuan Li and Caixia Chen. “Research on exercise fatigue estimation method of Pilates rehabilitation based on ECG and sEMG feature fusion”. In: *BMC Medical Informatics and Decision Making* 22.1 (Mar. 2022). DOI: 10.1186/s12911-022-01808-7.

[226]Huang Li, Shiaofen Fang, Snehasis Mukhopadhyay, Andrew J. Saykin, and Li Shen. “Interactive Machine Learning by Visualization: A Small Data Solution”. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2018. DOI: 10.1109/bigdata.2018.8621952.

[227]Jian Li, Yong Liu, Rong Yin, et al. “Multi-class learning: from theory to algorithm”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 1593–1602.

[228]Muyuan Li. “A Practical Significant Technic in Solving Overfitting: Regularization”. In: *Theoretical and Natural Science* 5.1 (May 2023), pp. 253–258. DOI: 10.54254/2753-8818/5/20230433.

[229]Na Li, Rui Zhou, Bharath Krishna, et al. “Non-invasive Techniques for Muscle Fatigue Monitoring: A Comprehensive Survey”. In: *ACM Computing Surveys* 56.9 (Apr. 2024), pp. 1–40. DOI: 10.1145/3648679.

[230]Shaojie Li and Yong Liu. “Sharper Generalization Bounds for Clustering”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 6392–6402.

[231]Xiali Li, Songting Deng, Song Wang, Zhengyu Lv, and Licheng Wu. “Review of Small Data Learning Methods”. In: *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, July 2018. DOI: 10.1109/compsac.2018.10212.

[232]Felipe Tomazelli Lima and Vinicius M.A. Souza. “A Large Comparison of Normalization Methods on Time Series”. In: *Big Data Research* 34 (Nov. 2023), p. 100407. DOI: 10.1016/j.bdr.2023.100407.

[233]Jonathan Feng-Shun Lin, Michelle Karg, and Dana Kulic. “Movement Primitive Segmentation for Human Motion Modeling: A Framework for Analysis”. In: *IEEE Transactions on Human-Machine Systems* 46.3 (June 2016), pp. 325–339. DOI: 10.1109/thms.2015.2493536.

[234]L. Litwin. “FIR and IIR digital filters”. In: *IEEE Potentials* 19.4 (2000), pp. 28–31. DOI: 10.1109/45.877863.

[235]Jingxuan Liu, Qing Tao, and Bin Wu. “Dynamic Muscle Fatigue State Recognition Based on Deep Learning Fusion Model”. In: *IEEE Access* 11 (2023), pp. 95079–95091. DOI: 10.1109/access.2023.3309741.

[236]Thiago Ribeiro Lopes, Hugo Maxwell Pereira, and Bruno Moreira Silva. “Perceived Exertion: Revisiting the History and Updating the Neurophysiology and the Practical Applications”. In: *International Journal of Environmental Research and Public Health* 19.21 (Nov. 2022), p. 14439. DOI: 10.3390/ijerph192114439.

[237]Miguel Bordallo Lopez, Carlos R. del-Blanco, and Narciso Garcia. “Detecting exercise-induced fatigue using thermal imaging and deep learning”. In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, Nov. 2017. DOI: 10.1109/ipta.2017.8310151.

[238]Hongyu Luo, Pierre-Alexandre Lee, Ieuan Clay, Martin Jaggi, and Valeria De Luca. “Assessment of Fatigue Using Wearable Sensors: A Pilot Study”. In: *Digital Biomarkers* 4.Suppl. 1 (Nov. 2020), pp. 59–72. DOI: 10.1159/000512166.

[239]Nan Ma. “Application of the Artificial Intelligence Sensor Principle in Athlete Training Status Testing System”. In: *2023 2nd International Conference on Artificial Intelligence and Computer Information Technology (AICIT)*. IEEE, Sept. 2023. DOI: 10.1109/aicit59054.2023.10277823.

[240]Wenhui Ma and Bin Guo. “Construction of neural network model for exercise load monitoring based on yoga training data and rehabilitation therapy”. In: *Heliyon* 10.12 (June 2024), e32679. DOI: 10.1016/j.heliyon.2024.e32679.

[241]Artur Magiera, Robert Roczniok, Ewa Sadowska-Krępa, et al. “The Effect of Physical And Mental Stress on the Heart Rate, Cortisol and Lactate Concentrations in Rock Climbers”. In: *Journal of Human Kinetics* 65.1 (Dec. 2018), pp. 111–123. DOI: 10.2478/hukin-2018-0024.

[242]Zahra Sedighi Maman, Ying-Ju Chen, Amir Baghdadi, et al. “A data analytic framework for physical fatigue management using wearable sensors”. In: *Expert Systems with Applications* 155 (Oct. 2020), p. 113405. DOI: 10.1016/j.eswa.2020.113405.

[243]Zahra Sedighi Maman, Mohammad Ali Alamdar Yazdi, Lora A. Cavuoto, and Fadel M. Megahed. “A data-driven approach to modeling physical fatigue in the workplace using wearable sensors”. In: *Applied Ergonomics* 65 (Nov. 2017), pp. 515–529. DOI: 10.1016/j.apergo.2017.02.001.

[244]Soumen Manna. “Small Sample Estimation of Classification Metrics”. In: *2022 Interdisciplinary Research in Technology and Management (IRTM)*. IEEE, Feb. 2022. DOI: 10.1109/irtm54583.2022.9791645.

[245]Nader Maray, Anne Hee Ngu, Jianyuan Ni, Minakshi Debnath, and Lu Wang. "Transfer Learning on Small Datasets for Improved Fall Detection". In: *Sensors* 23.3 (Jan. 2023), p. 1105. DOI: 10.3390/s23031105.

[246]Marco Marena, Neethan Ratnakumar, Rachel Jones, et al. "Predicting Metabolic Rate for Firefighting Activities with Worn Loads using a Heart Rate Sensor and Machine Learning". In: *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*. IEEE, Oct. 2023. DOI: 10.1109/bsn58485.2023.10331063.

[247]Luca Marotta, Bouke L. Scheltinga, Robbert van Middelaar, et al. "Accelerometer-Based Identification of Fatigue in the Lower Limbs during Cyclical Physical Exercise: A Systematic Review". In: *Sensors* 22.8 (Apr. 2022), p. 3008. DOI: 10.3390/s22083008.

[248]Neusa R. Adão Martins, Simon Annaheim, Christina M. Spengler, and René M. Rossi. "Fatigue Monitoring Through Wearables: A State-of-the-Art Review". In: *Frontiers in Physiology* 12 (Dec. 2021). DOI: 10.3389/fphys.2021.790292.

[249]Olivier Massin. "Defining Physical Efforts". In: (May 2022). DOI: 10.31234/osf.io/qmg5j.

[250]Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. *A Survey on Bias and Fairness in Machine Learning*. 2019. DOI: 10.48550/ARXIV.1908.09635.

[251]Harvey J. Miller. "The data avalanche is here. Shouldn't we be digging?" In: *Journal of Regional Science* 50.1 (Feb. 2010), pp. 181–201. DOI: 10.1111/j.1467-9787.2009.00641.x.

[252]Varun Mishra, Sougata Sen, Grace Chen, et al. "Evaluating the Reproducibility of Physiological Stress Detection Models". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.4 (Dec. 2020), pp. 1–29. DOI: 10.1145/3432220.

[253]Mehdi Mohammadi, Bijan Raahemi, Ahmad Akbari, and Babak Nassersharif. "Class dependent feature transformation for intrusion detection systems. Iranian Conference on Electrical Engineering". In: Parallel als Druckausg. erschienen. Piscataway, NJ: 19th Iranian Conference on Electrical Engineering (ICEE), 2011.

[254]Mehryar Mohri. *Foundations of machine learning*. Ed. by Afshin Rostamizadeh and Ameet Talwalkar. Second edition. Adaptive computation and machine learning. Literaturverzeichnis: Seite 461-474. Cambridge, Massachusetts: The MIT Press, 2018. 486 pp.

[255]Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. "RecoFit". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2014. DOI: 10.1145/2556288.2557116.

[256]Siqi Mu, Shiwei Liao, Kuan Tao, and Yanfei Shen. "Intelligent fatigue detection based on hierarchical multi-scale ECG representations and HRV measures". In: *Biomedical Signal Processing and Control* 92 (June 2024), p. 106127. DOI: 10.1016/j.bspc.2024.106127.

[257] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. “Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization”. In: *Advances in Computational Mathematics* 25.1–3 (July 2006), pp. 161–193. DOI: 10.1007/s10444-004-7634-z.

[258] Alhassan Mumuni and Fuseini Mumuni. “Data augmentation: A comprehensive survey of modern approaches”. In: *Array* 16 (Dec. 2022), p. 100258. DOI: 10.1016/j.array.2022.100258.

[259] Kevin Murphy. *Machine Learning - A Probabilistic Perspective*. Adaptive Computation and Machine Learning. Description based upon print version of record. Cambridge: MIT Press, 2012. 1098 pp.

[260] A. I. Naimi and D. J. Westreich. “Big Data: A Revolution That Will Transform How We Live, Work, and Think”. In: *American Journal of Epidemiology* 179.9 (Apr. 2014), pp. 1143–1144. DOI: 10.1093/aje/kwu085.

[261] Farnad Nasirzadeh, Mostafa Mir, Sadiq Hussain, et al. “Physical Fatigue Detection Using Entropy Analysis of Heart Rate Signals”. In: *Sustainability* 12.7 (Mar. 2020), p. 2714. DOI: 10.3390/su12072714.

[262] Alejandro Newell, Kaiyu Yang, and Jia Deng. *Stacked Hourglass Networks for Human Pose Estimation*. 2016. DOI: 10.48550/ARXIV.1603.06937.

[263] Timothy David Noakes. “Challenging beliefs: ex Africa semper aliquid novi”. In: *Medicine & Science in Sports & Exercise* 29.5 (May 1997), pp. 571–590. DOI: 10.1097/00005768-199705000-00001.

[264] Barry S. Oken, Irina Chamine, and Wayne Wakeland. “A systems approach to stress, stressors and resilience in humans”. In: *Behavioural Brain Research* 282 (Apr. 2015), pp. 144–154. DOI: 10.1016/j.bbr.2014.12.047.

[265] Michalis Papakostas, Varun Kanal, Maher Abujelala, Konstantinos Tsiakas, and Fillia Makedon. “Physical fatigue detection through EMG wearables and subjective user reports”. In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, June 2019. DOI: 10.1145/3316782.3322772.

[266] Michael Parker. *Digital Signal Processing 101. Everything You Need to Know to Get Started*. Second edition. Includes bibliographical references at the end of each chapters and index. Oxford, United Kingdom: Newnes is an imprint of Elsevier, 2017. 11 pp.

[267] Kurt Partridge and Philippe Golle. “On using existing time-use study data for ubiquitous computing applications”. In: *Proceedings of the 10th international conference on Ubiquitous computing*. UbiComp08. ACM, Sept. 2008. DOI: 10.1145/1409635.1409655.

[268] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. “Investigating statistical machine learning as a tool for software development”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. ACM, Apr. 2008. DOI: 10.1145/1357054.1357160.

[269]Nathalie Pattyn, Jeroen Van Cutsem, Emilie Dessy, and Olivier Mairesse. “Bridging Exercise Science, Cognitive Psychology, and Medical Practice: Is "Cognitive Fatigue" a Remake of "The Emperor’s New Clothes"?” In: *Frontiers in Psychology* 9 (Sept. 2018). DOI: 10.3389/fpsyg.2018.01246.

[270]Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. “Effective Techniques for Multimodal Data Fusion: A Comparative Analysis”. In: *Sensors* 23.5 (Feb. 2023), p. 2381. DOI: 10.3390/s23052381.

[271]Roger Peng. “The Reproducibility Crisis in Science: A Statistical Counterattack”. In: *Significance* 12.3 (June 2015), pp. 30–32. DOI: 10.1111/j.1740-9713.2015.00827.x.

[272]Ana Pereira, Duarte Folgado, Ricardo Cotrim, and Inês Sousa. “Physiotherapy Exercises Evaluation using a Combined Approach based on sEMG and Wearable Inertial Sensors”. In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2019. DOI: 10.5220/0007391300730082.

[273]João Gonçalo Pereira and Joaquim Gonçalves. “Human Activity Recognition: A review”. In: *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, June 2022. DOI: 10.1109/isdfs55398.2022.9800781.

[274]Igor Pernek, Gregorij Kurillo, Gregor Stiglic, and Ruzena Bajcsy. “Recognizing the intensity of strength training exercises with wearable sensors”. In: *Journal of Biomedical Informatics* 58 (Dec. 2015), pp. 145–155. DOI: 10.1016/j.jbi.2015.09.020.

[275]David Perpetuini, Damiano Formenti, Daniela Cardone, et al. “Can Data-Driven Supervised Machine Learning Approaches Applied to Infrared Thermal Imaging Data Estimate Muscular Activity and Fatigue?” In: *Sensors* 23.2 (Jan. 2023), p. 832. DOI: 10.3390/s23020832.

[276]Ross O. Phillips. “A review of definitions of fatigue – And a step towards a whole definition”. In: *Transportation Research Part F: Traffic Psychology and Behaviour* 29 (Feb. 2015), pp. 48–56. DOI: 10.1016/j.trf.2015.01.003.

[277]Swetha Pinjerla, Srinivasa Rao S, and Chandrasekhar Reddy P. “Sampling Rate Conversion Techniques- A Review”. In: *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*. IEEE, Feb. 2022. DOI: 10.1109/icrtcst54752.2022.9781914.

[278]Cristina-Ioana Pirscoveanu and Anderson Souza Oliveira. “Prediction of instantaneous perceived effort during outdoor running using accelerometry and machine learning”. In: *European Journal of Applied Physiology* 124.3 (Sept. 2023), pp. 963–973. DOI: 10.1007/s00421-023-05322-0.

[279]Gede Angga Pradipta, Retantyo Wardoyo, Aina Musdholifah, I Nyoman Hariyasa Sanjaya, and Muhammad Ismail. “SMOTE for Handling Imbalanced Data Problem: A Review”. In: *2021 Sixth International Conference on Informatics and Computing (ICIC)*. IEEE, Nov. 2021. DOI: 10.1109/icic54025.2021.9632912.

[280]D. V. Pravin, A. J. Ragavkumar, S. Abinesh, and G. Kavitha. “Extraction, Processing and Analysis of Surface Electromyogram Signal and Detection of Muscle Fatigue Using Machine Learning Methods”. In: *2023 International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*. IEEE, Mar. 2023. DOI: 10.1109/icbsii58188.2023.10181085.

[281]Nicholas Pudjihartono, Tayaza Fadason, Andreas W. Kempa-Liehr, and Justin M. O’Sullivan. “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction”. In: *Frontiers in Bioinformatics* 2 (June 2022). DOI: 10.3389/fbinf.2022.927312.

[282]Lisandro Puglisi, Roque Saltaren, and Cecilia Garcia Cena. “On the Velocity and Acceleration Estimation from Discrete Time-Position Sensors”. In: *Control Engineering and Applied Informatics* 17 (Oct. 2015), pp. 30–40.

[283]Guo-Jun Qi and Jiebo Luo. “Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.4 (Apr. 2022), pp. 2168–2187. DOI: 10.1109/tpami.2020.3031898.

[284]Maria Rauschenberger and Ricardo Baeza-Yates. “How to Handle Health-Related Small Imbalanced Data in Machine Learning?” In: *i-com* 19.3 (Dec. 2020), pp. 215–226. DOI: 10.1515/icom-2020-0018.

[285]D.A. Reid, S. Samangooei, C. Chen, M.S. Nixon, and A. Ross. “Soft Biometrics for Surveillance: An Overview”. In: *Handbook of Statistics - Machine Learning: Theory and Applications*. Elsevier, 2013, pp. 327–352. DOI: 10.1016/b978-0-444-53859-8.00013-8.

[286]Pavel Ripka, ed. *Magnetic sensors and magnetometers*. Second edition. Includes bibliographical references and index. Norwood, MA: Artech House, 2021. 1 p.

[287]Miao Rong, Dunwei Gong, and Xiaozhi Gao. “Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends”. In: *IEEE Access* 7 (2019), pp. 19709–19725. DOI: 10.1109/access.2019.2894366.

[288]Milad Sadat-Mohammadi, Shahrad Shakerian, Yizhi Liu, Somayeh Asadi, and Houtan Jebelli. “Non-invasive physical demand assessment using wearable respiration sensor and random forest classifier”. In: *Journal of Building Engineering* 44 (Dec. 2021), p. 103279. DOI: 10.1016/j.jobe.2021.103279.

[289]İbrahim Halil Şahin and Ahmet Sanioğlu. “The examination of the relationship between body composition and acceleration”. In: *Turkish Journal of Kinesiology* 9.2 (June 2023), pp. 106–114. DOI: 10.31459/turkjkin.1295059.

[290]Abderahim Salhi, Althea C. Henslee, James Ross, Joseph Jabour, and Ian Dettwiller. “Data Preprocessing Using AutoML: A Survey”. In: *2023 Congress in Computer Science, Computer Engineering & Applied Computing (CSCE)*. IEEE, July 2023. DOI: 10.1109/csce60160.2023.00265.

[291]A. L. Samuel. “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM Journal of Research and Development* 3.3 (July 1959), pp. 210–229. DOI: 10.1147/rd.33.0210.

[292]Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: (2018). DOI: 10.48550/ARXIV.1801.04381.

[293]Zahra Alizadeh Sani, Mohammad Tayarani Darbandy, Mozhdeh Rostamnezhad, et al. “A new approach to detect the physical fatigue utilizing heart rate signals”. In: *Research in Cardiovascular Medicine* 9.1 (2020), p. 23. DOI: 10.4103/rcm.rcm_8_20.

[294]Yutaka Sasaki. “The truth of the F-measure”. In: *Teach Tutor Mater* (Jan. 2007).

[295]Mark Saunders. *Research methods for business students*. Ed. by Adrian Thornhill and Philip Lewis. Harlow, United Kingdom, 2019.

[296]Amir Masoud Sefidian. *Understanding Micro, Macro, and Weighted Averages for Scikit-Learn metrics in multi-class classification with example*. https://iamirmasoud.com/2022/06/19/understanding-micro-macro-and-weighted-averages-for-scikit-learn-metrics-in-multi-class-classification-with-example/. Sept. 2024.

[297]Martin Seiffert, Flavio Holstein, Rainer Schlosser, and Jochen Schiller. “Next Generation Cooperative Wearables: Generalized Activity Assessment Computed Fully Distributed Within a Wireless Body Area Network”. In: *IEEE Access* 5 (2017), pp. 16793–16807. DOI: 10.1109/access.2017.2749005.

[298]H. Selye. “The stress of life”. In: *New York* (1956).

[299]Dhruv R. Seshadri, Ryan T. Li, James E. Voos, et al. “Wearable sensors for monitoring the physiological and biochemical profile of the athlete”. In: *npj Digital Medicine* 2.1 (July 2019). DOI: 10.1038/s41746-019-0150-9.

[300]Mert Sevil, Mudassir Rashid, Mohammad Reza Askari, et al. “Detection and Characterization of Physical Activity and Psychological Stress from Wristband Data”. In: *Signals* 1.2 (Dec. 2020), pp. 188–208. DOI: 10.3390/signals1020011.

[301]Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. “Representation Bias in Data: A Survey on Identification and Resolution Techniques”. In: *ACM Computing Surveys* 55.13s (July 2023), pp. 1–39. DOI: 10.1145/3588433.

[302]Sina Shahmoradi, Alireza Zare, and Saeed Behzadipour. “Fatigue Status Recognition in a Post-Stroke Rehabilitation Exercise with sEMG Signal”. In: *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*. Vol. 82. IEEE, Nov. 2017, pp. 1–5. DOI: 10.1109/icbme.2017.8430264.

[303]Torgyn Shaikhina, Dave Lowe, Sunil Daga, et al. “Machine Learning for Predictive Modelling based on Small Data in Biomedical Engineering”. In: *IFAC-PapersOnLine* 48.20 (2015), pp. 469–474. DOI: 10.1016/j.ifacol.2015.10.185.

[304]Harsh Sharma and Anushika Gosain. “Oversampling Methods to Handle the Class Imbalance Problem: A Review”. In: *Soft Computing and Its Engineering Applications*. Springer Nature Switzerland, 2023, pp. 96–110. DOI: 10.1007/978-3-031-27609-5_8.

[305] Deyao Shen, Xuyuan Tao, Vladan Koncar, and Jianping Wang. “A Review of Intelligent Garment System for Bioelectric Monitoring During Long-Lasting Intensive Sports”. In: *IEEE Access* 11 (2023), pp. 111358–111377. DOI: 10.1109/access.2023.3322925.

[306] Song Shi, Ziping Cao, Hengheng Li, et al. “Recognition System of Human Fatigue State Based on Hip Gait Information in Gait Patterns”. In: *Electronics* 11.21 (Oct. 2022), p. 3514. DOI: 10.3390/electronics11213514.

[307] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. “Ecological Momentary Assessment”. In: *Annual Review of Clinical Psychology* 4.1 (Apr. 2008), pp. 1–32. DOI: 10.1146/annurev.clinpsy.3.022806.091415.

[308] Pannaga Shivaswamy and Ashok Chandrashekar. “Bias-Variance Decomposition for Ranking”. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. WSDM ’21. ACM, Mar. 2021. DOI: 10.1145/3437963.3441772.

[309] Connor Shorten and Taghi M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. In: *Journal of Big Data* 6.1 (July 2019). DOI: 10.1186/s40537-019-0197-0.

[310] T.S. Sindhu, N. Kumaratharan, and P. Anandan. “A Review on Optimization Algorithms for Feature Selection”. In: *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. IEEE, June 2023. DOI: 10.1109/icscss57650.2023.10169444.

[311] Simon Skau, Kristoffer Sundberg, and Hans-Georg Kuhn. “A Proposal for a Unifying Set of Definitions of Fatigue”. In: *Frontiers in Psychology* 12 (Oct. 2021). DOI: 10.3389/fpsyg.2021.739764.

[312] Paul E. Smaldino and Richard McElreath. “The natural selection of bad science”. In: *Royal Society Open Science* 3.9 (Sept. 2016), p. 160384. DOI: 10.1098/rsos.160384.

[313] Aref Smiley, Te-Yi Tsai, Elena Zakashansky, et al. “Exercise Exertion Levels Prediction Based on Real-Time Wearable Physiological Signal Monitoring”. In: *Healthcare Transformation with Informatics and Artificial Intelligence*. IOS Press, June 2023. DOI: 10.3233/shti230454.

[314] Steven W. Smith. *The scientist and engineer’s guide to digital signal processing*. California Technical Pub., 1999, p. 650.

[315] Vimalraj S Spelmen and R Porkodi. “A Review on Handling Imbalanced Data”. In: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE, Mar. 2018. DOI: 10.1109/icctct.2018.8551020.

[316] Josh Starmer. *The StatQuest Illustrated Guide To Machine Learning*. 305 pages. Independently published, May 2022.

[317] Matthew A. Stults-Kolehmainen and Rajita Sinha. “The Effects of Stress on Physical Activity and Exercise”. In: *Sports Medicine* 44.1 (Sept. 2013), pp. 81–121. DOI: 10.1007/s40279-013-0090-5.

[318] Yunxiang Su, Yikun Gong, and Shaoxu Song. "Time Series Data Validity". In: *Proceedings of the ACM on Management of Data* 1.1 (May 2023), pp. 1–26. DOI: 10.1145/3588939.

[319] Abdulhamit Subasi and M. Kemal Kiymik. "Muscle Fatigue Detection in EMG Using Time–Frequency Methods, ICA and Neural Networks". In: *Journal of Medical Systems* 34.4 (Apr. 2009), pp. 777–785. DOI: 10.1007/s10916-009-9292-7.

[320] Jiaqi Sun, Guangda Liu, Yubing Sun, et al. "Application of Surface Electromyography in Exercise Fatigue: A Review". In: *Frontiers in Systems Neuroscience* 16 (Aug. 2022). DOI: 10.3389/fnsys.2022.893275.

[321] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. "Classification of Imbalanced Data: a Review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.04 (June 2009), pp. 687–719. DOI: 10.1142/s0218001409007326.

[322] Amal Tawakuli, Bastian Havers, Vincenzo Gulisano, Daniel Kaiser, and Thomas Engel. "Survey:Time-series data preprocessing: A survey and an empirical analysis". In: *Journal of Engineering Research* (Mar. 2024). DOI: 10.1016/j.jer.2024.02.018.

[323] Portia E Taylor, Gustavo J M Almeida, Takeo Kanade, and Jessica K Hodgins. "Classifying human motion quality for knee osteoarthritis using accelerometers". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, Aug. 2010. DOI: 10.1109/iembs.2010.5627665.

[324] Mamello Thinyane. "Investigating an Architectural Framework for Small Data Platforms". In: *Data for societal challenges-17th European Conference on Digital Government (ECDG 2017)* (2017), pp. 220–227.

[325] Catherine Tong, Shyam A. Tailor, and Nicholas D. Lane. "Are Accelerometers for Activity Recognition a Dead-end?" In: *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*. ACM, Mar. 2020. DOI: 10.1145/3376897.3377867.

[326] José Francisco Tornero-Aguilera, Jorge Jimenez-Morcillo, Alejandro Rubio-Zarapuz, and Vicente J. Clemente-Suárez. "Central and Peripheral Fatigue in Physical Exercise Explained: A Narrative Review". In: *International Journal of Environmental Research and Public Health* 19.7 (Mar. 2022), p. 3909. DOI: 10.3390/ijerph19073909.

[327] Andreas Triantafyllopoulos, Sandra Ottl, Alexander Gebhard, et al. *Fatigue Prediction in Outdoor Running Conditions using Audio Data*. 2022. DOI: 10.48550/ARXIV.2205.04343.

[328] John Trimpop, Hannes Schenk, Gerald Bieber, Friedrich Lämmel, and Paul Burggraf. "Smartwatch based Respiratory Rate and Breathing Pattern Recognition in an End-consumer Environment". In: *Proceedings of the 4th International Workshop on Sensor-based Activity Recognition and Interaction*. ACM, Sept. 2017. DOI: 10.1145/3134230.3134235.

[329] Iryna Trygub, Johanna Ahlf, Martina Campanale, André Jeworutzki, and Jan Schwarzer. “NeckWatcher: A Real-time Monitoring Tool for the Assessment of the Neck Posture”. In: *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’23. ACM, July 2023. DOI: 10.1145/3594806.3596529.

[330] Jiaobing Tu and Wei Gao. “Ethical Considerations of Wearable Technologies in Human Research”. In: *Advanced Healthcare Materials* 10.17 (Apr. 2021). DOI: 10.1002/adhm.202100127.

[331] Waleed Umer, Heng Li, Yu Yantao, et al. “Physical exertion modeling for construction tasks using combined cardiorespiratory and thermoregulatory measures”. In: *Automation in Construction* 112 (Apr. 2020), p. 103079. DOI: 10.1016/j.autcon.2020.103079.

[332] Waleed Umer, Yantao Yu, Maxwell Fordjour Antwi-Afari, et al. “Heart rate variability based physical exertion monitoring for manual material handling tasks”. In: *International Journal of Industrial Ergonomics* 89 (May 2022), p. 103301. DOI: 10.1016/j.ergon.2022.103301.

[333] Padhraic Smyth Usama Fayyad Gregory Piatetsky-Shapiro. “From Data Mining to Knowledge Discovery in Databases”. In: *AI Magazine* 17.3 (Mar. 1996), pp. 37–54. DOI: 10.1609/aimag.v17i3.1230.

[334] F. Valeriani, C. Protano, A. De Giorgi, et al. “Analysing features of home-based workout during COVID-19 pandemic: a systematic review”. In: *Public Health* 222 (Sept. 2023), pp. 100–114. DOI: 10.1016/j.puhe.2023.06.040.

[335] Elli Valla, Ain-Joonas Toose, Sven Nõmm, and Aaro Toomela. “Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests”. In: *International Journal of Medical Informatics* 177 (Sept. 2023), p. 105152. DOI: 10.1016/j.ijmedinf.2023.105152.

[336] Gilles Vandewiele, Youri Geurkink, Maarten Lievens, et al. “Enabling training personalization by predicting the session rate of perceived exertion (sRPE)”. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. 2017.

[337] V. N. Vapnik and A. Ya. Chervonenkis. “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities”. In: *Measures of Complexity*. Springer International Publishing, 2015, pp. 11–30. DOI: 10.1007/978-3-319-21852-6_3.

[338] Andreas Venhorst, Dominic Micklewright, and Timothy D. Noakes. “Perceived Fatigability: Utility of a Three-Dimensional Dynamical Systems Framework to Better Understand the Psychophysiological Regulation of Goal-Directed Exercise Behaviour”. In: *Sports Medicine* 48.11 (Sept. 2018), pp. 2479–2495. DOI: 10.1007/s40279-018-0986-1.

[339] Michalis Vrigkas, Christophoros Nikou, and Ioannis A. Kakadiaris. “A Review of Human Activity Recognition Methods”. In: *Frontiers in Robotics and AI* 2 (Nov. 2015). DOI: 10.3389/frobt.2015.00028.

[340] Guido Walz. *Interpolation Von Daten und Funktionen. Klartext Für Nichtmathematiker*. Essentials Ser. Description based on publisher supplied metadata and other sources. Wiesbaden: Springer Fachmedien Wiesbaden GmbH, 2020. 161 pp.

[341] Bin Wang and Dongzhi He. "Prediction method of running fatigue based on depth image". In: *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. IEEE, June 2021. DOI: 10.1109/imcec51613.2021.9482204.

[342] Chuanling Wang, Xi Wang, Qiao Li, and Xiaoming Tao. "Recognizing and predicting muscular fatigue of biceps brachii in motion with novel fabric strain sensors based on machine learning". In: *Biomedical Signal Processing and Control* 96 (Oct. 2024), p. 106647. DOI: 10.1016/j.bspc.2024.106647.

[343] Guodong Wang, Xiaokun Mao, Qiuxia Zhang, and Aming Lu. "Fatigue Detection in Running with Inertial Measurement Unit and Machine Learning". In: *2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB)*. IEEE, May 2022. DOI: 10.1109/icbcb55259.2022.9802471.

[344] Junhong Wang, Shaoming Sun, and Yining Sun. "A Muscle Fatigue Classification Model Based on LSTM and Improved Wavelet Packet Threshold". In: *Sensors* 21.19 (Sept. 2021), p. 6369. DOI: 10.3390/s21196369.

[345] Junhong Wang, Yining Sun, and Shaoming Sun. "Recognition of Muscle Fatigue Status Based on Improved Wavelet Threshold and CNN-SVM". In: *IEEE Access* 8 (2020), pp. 207914–207922. DOI: 10.1109/access.2020.3038422.

[346] Pingan Wang, Ju-Seok Nam, and Xiongzhe Han. "Development of a comprehensive fatigue detection model for beekeeping activities based on deep learning and EEG signals". In: *Computers and Electronics in Agriculture* 225 (Oct. 2024), p. 109265. DOI: 10.1016/j.compag.2024.109265.

[347] Chathura Wanigasekara, Akshya Swain, Sing Kiong Nguang, and B. Gangadhara Prusty. "Neural Network Based Inverse System Identification from Small Data Sets". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2019. DOI: 10.1109/ijcnn.2019.8851722.

[348] Anna Wexler and Emily Largent. "Ethical considerations for researchers developing and testing minimal-risk devices". In: *Nature Communications* 14.1 (Apr. 2023). DOI: 10.1038/s41467-023-38068-6.

[349] R.M. White. "A Sensor Classification Scheme". In: *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control* 34.2 (Mar. 1987), pp. 124–126. DOI: 10.1109/t-uffc.1987.26922.

[350] David A. Winter. *Biomechanics and Motor Control of Human Movement*. Wiley, Sept. 2009. DOI: 10.1002/9780470549148.

[351] Johnny Chun Yiu Wong, Jun Wang, Eugene Yujun Fu, Hong Va Leong, and Grace Ngai. "Activity Recognition and Stress Detection via Wristband". In: *Proceedings of the 17th International Conference on Advances in Mobile Computing & Multimedia*. ACM, Dec. 2019. DOI: 10.1145/3365921.3365950.

[352] Iain H. Woodhouse. “On ‘ground’ truth and why we should abandon the term”. In: *Journal of Applied Remote Sensing* 15.04 (Nov. 2021). DOI: 10.1117/1.jrs.15.041501.

[353] Ming-Yen Wu, Chi-Hua Chen, and Chi-Chun Lo. *An Exercise Fatigue Detection Model Based on Machine Learning Methods*. 2018. DOI: 10.48550/ARXIV.1803.07952.

[354] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. “Small data machine learning in materials science”. In: *npj Computational Materials* 9.1 (Mar. 2023). DOI: 10.1038/s41524-023-01000-z.

[355] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. “Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models”. In: *Proceedings of the 2018 Designing Interactive Systems Conference*. DIS ’18. ACM, June 2018. DOI: 10.1145/3196709.3196729.

[356] Tao Yang and Vojislav Kecman. “Adaptive local hyperplane algorithm for learning small medical data sets”. In: *Expert Systems* 26.4 (Sept. 2009), pp. 355–359. DOI: 10.1111/j.1468-0394.2009.00494.x.

[357] Zhongwan Yang and Huijie Ren. “Feature Extraction and Simulation of EEG Signals During Exercise-Induced Fatigue”. In: *IEEE Access* 7 (2019), pp. 46389–46398. DOI: 10.1109/access.2019.2909035.

[358] Hongyan Yao. “Prediction of sports fatigue degree based on spectral sensors and machine learning algorithms”. In: *Optical and Quantum Electronics* 56.4 (Feb. 2024). DOI: 10.1007/s11082-024-06531-3.

[359] Gal Yarin. “Uncertainty in Deep Learning”. PhD thesis. University of Cambridge, Sept. 2016.

[360] Merve Nur Yasar, Marco Sica, Brendan O’Flynn, Salvatore Tedesco, and Matteo Menolotto. “A dataset for fatigue estimation during shoulder internal and external rotation movements using wearables”. In: *Scientific Data* 11.1 (Apr. 2024). DOI: 10.1038/s41597-024-03254-8.

[361] Runpeng Yu, Hong Zhu, Kaican Li, et al. *Regularization Penalty Optimization for Addressing Data Quality Variance in OoD Algorithms*. 2022. DOI: 10.48550/ARXIV.2206.05749.

[362] Abdulaziz Zamkah, Terence Hui, Simon Andrews, et al. “Identification of Suitable Biomarkers for Stress and Emotion Detection for Future Personal Affective Wearable Sensors”. In: *Biosensors* 10.4 (Apr. 2020), p. 40. DOI: 10.3390/bios10040040.

[363] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (Feb. 2021), pp. 107–115. DOI: 10.1145/3446776.

[364] Fan Zhang and Feng Wang. “Exercise Fatigue Detection Algorithm Based on Video Image Information Extraction”. In: *IEEE Access* 8 (2020), pp. 199696–199709. DOI: 10.1109/access.2020.3023648.

[365] Guoxin Zhang, Tommy Tung-Ho Hong, Li Li, and Ming Zhang. "Automatic Detection of Fatigued Gait Patterns in Older Adults: An Intelligent Portable Device Integrating Force and Inertial Measurements with Machine Learning". In: *Annals of Biomedical Engineering* (Aug. 2024). DOI: 10.1007/s10439-024-03603-z.

[366] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, et al. "A Comprehensive Survey of Vision-Based Human Action Recognition Methods". In: *Sensors* 19.5 (Feb. 2019), p. 1005. DOI: 10.3390/s19051005.

[367] Jian Zhang, Thurmon E. Lockhart, and Rahul Soangra. "Classifying Lower Extremity Muscle Fatigue During Walking Using Machine Learning and Inertial Sensors". In: *Annals of Biomedical Engineering* 42.3 (Oct. 2013), pp. 600–612. DOI: 10.1007/s10439-013-0917-0.

[368] Wentong Zhang, Caixia Su, and Chuan He. "Rehabilitation Exercise Recognition and Evaluation Based on Smart Sensors With Deep Learning Framework". In: *IEEE Access* 8 (2020), pp. 77561–77571. DOI: 10.1109/access.2020.2989128.

[369] Ying Zhang and Chen Ling. "A strategy to apply machine learning to small datasets in materials science". In: *npj Computational Materials* 4.1 (May 2018). DOI: 10.1038/s41524-018-0081-z.

[370] Yongqing Zhang, Siyu Chen, Wenpeng Cao, et al. "MFFNet: Multi-dimensional Feature Fusion Network based on attention mechanism for sEMG analysis to detect muscle fatigue". In: *Expert Systems with Applications* 185 (Dec. 2021), p. 115639. DOI: 10.1016/j.eswa.2021.115639.

[371] Zimu Zhang and Xiujian Zhang. "A Review of Research on Generalization Error Analysis of Deep Learning Models". In: *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, Nov. 2023. DOI: 10.1109/icicml60161.2023.10424837.

[372] Yantong Zheng and Victoria Stodden. "The Idealized Machine Learning Pipeline (IMLP) for Advancing Reproducibility in Machine Learning". In: *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*. ACM REP '24. ACM, June 2024. DOI: 10.1145/3641525.3663630.

[373] Shaoxuan Zhou. "An Analysis of The Small Sample Datasets Based on Machine Learning". In: *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*. EITCE 2022. ACM, Oct. 2022. DOI: 10.1145/3573428.3573720.

[374] Haiyan Zhu, Yuelong Ji, Baiyang Wang, and Yuyun Kang. "Exercise fatigue diagnosis method based on short-time Fourier transform and convolutional neural network". In: *Frontiers in Physiology* 13 (Aug. 2022). DOI: 10.3389/fphys.2022.965974.

[375] Min Zhu, Jing Xia, Xiaoqing Jin, et al. "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data". In: *IEEE Access* 6 (2018), pp. 4641–4652. DOI: 10.1109/access.2018.2789428.

[376] Honglei Zhuang, Xuanhui Wang, Michael Bendersky, and Marc Najork. “Feature Transformation for Neural Ranking Models”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’20. ACM, July 2020. DOI: 10.1145/3397271.3401333.

[377] Aniket Zinzuwadia and Jagmeet P Singh. “Wearable devices—addressing bias and inequity”. In: *The Lancet Digital Health* 4.12 (Dec. 2022), e856–e857. DOI: 10.1016/s2589-7500(22)00194-7.

Contributions

A

The following Table A.1 illustrates the contributions of this thesis compared to related works. The column "Generalisation & Evaluation" highlights works that have contributed an overview or analysis of ML generalisation and evaluation methods for fatigue detection with small data. The column "Fatigue Framework" highlights works that have contributed a framework for exercise fatigue detection with ML and small data. The column "2D PE Fatigue Detection" highlights works that have used PE based on 2D cameras for fatigue detection. The column "IMU vs. PE Comparison" highlights works that have compared PE and IMU sensors for exercise fatigue detection. The column "Increasing n" highlights works that have investigated ML performance for an increasing number of subjects. The column "Small Data Augmentation" highlights related works that have applied data augmentation techniques to small data. The column "Multiple Evaluation Types" highlights related works that have used multiple evaluation types for ML.

Tab. A.1.: Contributions of this thesis compared to the related works.

Authors	Year	n	Generalisation & Evaluation	Fatigue Framework	2D PE Fatigue Detection	IMU vs PE Comparison	Increasing n	Small Data Augmentation	Multiple Evaluation Types
Albert and Arnrich [9]	2024	16	-	-	-	-	-	x	-
Gan et al. [126]	2024	16	-	-	-	-	-	-	N/A
Huang et al. [161]	2024	7	-	-	-	-	-	N/A	-
Ma and Guo [240]	2024	30	-	-	-	-	-	N/A	N/A
Mu et al. [256]	2024	32	-	-	-	-	-	-	-
Wang et al. [342]	2024	32	-	-	-	-	-	-	-
Wang et al. [346]	2024	5	-	-	-	-	-	x	-
Yao [358]	2024	20	-	-	-	-	-	-	N/A
Zhang et al. [365]	2024	18	-	-	-	-	-	-	-
Adapa et al. [1]	2023	11	-	-	-	-	-	-	-
Antwi-Afari et al. [20]	2023	10	-	-	-	-	-	-	-

Table A.1 continued from the previous page

Authors	Year	n	Generalisation & Eval. Types	Fatigue Framework	2D PE Fatigue Detection	IMU vs PE Comparison	Increasing n	Small Data Augmentation	Multiple Eval. Types
Anwer et al. [21]	2023	15	-	-	-	-	-	-	-
Biró et al. [45]	2023	9	-	-	-	-	-	x	-
Biró et al. [44]	2023	19	-	-	-	-	-	-	N/A
Bouteraa et al. [52]	2023	57	-	-	-	-	-	-	N/A
Cañellas et al. [63]	2023	80	-	-	-	-	-	x	N/A
Concha-Pérez et al. [83]	2023	30	-	-	-	-	-	-	x
Dang et al. [92]	2023	10	-	-	-	-	-	-	-
De Vito et al. [97]	2023	1	-	-	-	-	-	-	-
Dimmick et al. [102]	2023	16	-	-	-	-	-	-	x
Feng et al. [116]	2023	25	-	-	-	-	-	-	-
Kathirgamanathan et al. [191]	2023	19	-	-	-	-	-	-	x
Liu et al. [235]	2023	20	-	-	-	-	-	-	-
Marena et al. [246]	2023	5	-	-	-	-	-	-	-
Perpetuini et al. [275]	2023	10	-	-	-	-	-	-	-
Pirscoveanu and Oliveira [278]	2023	43	-	-	-	-	-	-	x
Pravin et al. [280]	2023	N/A	-	-	-	-	-	-	N/A
Smiley et al. [313]	2023	10	-	-	-	-	-	-	-
Valla et al. [335]	2023	41	-	-	-	-	-	-	-
Albert et al. [10]	2022	12	-	-	-	-	-	-	-
Bustos et al. [58]	2022	24	-	-	-	-	-	-	x
Jaiswal et al. [168]	2022	32	-	-	-	-	-	-	-
Umer et al. [332]	2022	10	-	-	-	-	-	-	-
Cheah et al. [69]	2022	4	-	-	-	-	-	-	-
Escobar-Linero et al. [112]	2022	7	-	-	-	-	-	-	-
Guo et al. [139]	2022	10	-	-	-	-	-	-	-
Jiang et al. [176]	2022	12	-	-	-	-	-	-	-
Li and Chen [225]	2022	20	-	-	-	-	-	-	-
Shi et al. [306]	2022	10	-	-	-	-	-	-	-
Triantafyllopoulos et al. [327]	2022	48	-	-	-	-	-	-	-
Wang et al. [343]	2022	19	-	-	-	-	-	-	-
Zhu et al. [374]	2022	24	-	-	-	-	-	-	-
Chen et al. [71]	2021	40	-	-	-	-	-	-	-
Chen et al. [70]	2021	47	-	-	-	-	-	-	-
K et al. [181]	2021	58	-	-	-	-	-	-	-
Sadat-Mohammadi et al. [288]	2021	15	-	-	-	-	-	-	-
Wang et al. [344]	2021	20	-	-	-	-	-	-	-
Zhang et al. [370]	2021	10	-	-	-	-	-	-	x
Aguirre et al. [4]	2021	60	-	-	-	-	-	-	-
Balaskas and Siozios [28]	2021	14	-	-	-	-	-	-	N/A
Chalitsios et al. [67]	2021	13	-	-	-	-	-	-	-
Chen et al. [72]	2021	10	-	-	-	-	-	-	-
Elshafei et al. [108]	2021	20	-	-	-	-	-	-	x
Guan et al. [138]	2021	14	-	-	-	-	-	x	-
Jiang et al. [175]	2021	14	-	-	-	-	-	x	-
Karvekar et al. [189]	2021	24	-	-	-	-	-	-	-

Table A.1 continued from the previous page

Authors	Year	n	Generalisation & Eval. Types	Fatigue Framework	2D PE Fatigue Detection	IMU vs PE Comparison	Increasing n	Small Data Augmentation	Multiple Eval. Types
Kuschan and Krüger [212]	2021	9	-	-	-	-	-	-	N/A
Lambay et al. [216]	2021	24	-	-	-	-	-	-	-
Wang and He [341]	2021	12	-	-	-	-	-	-	x
Davidson et al. [95]	2020	12	-	-	-	-	-	x	-
Luo et al. [238]	2020	27	-	-	-	-	-	-	-
Umer et al. [331]	2020	10	-	-	-	-	-	-	x
Wang et al. [345]	2020	20	-	-	-	-	-	-	-
Guaitolini et al. [137]	2020	13	-	-	-	-	-	-	-
Maman et al. [242]	2020	15	-	-	-	-	-	-	-
Nasirzadeh et al. [261]	2020	8	-	-	-	-	-	-	-
Sani et al. [293]	2020	8	-	-	-	-	-	-	N/A
Zhang and Wang [364]	2020	20	-	-	-	-	-	-	-
Chowdhury et al. [77]	2019	22	-	-	-	-	-	-	-
Geurkink et al. [129]	2019	46	-	-	-	-	-	-	-
Jebelli et al. [171]	2019	10	-	-	-	-	-	-	-
Karvekar et al. [188]	2019	24	-	-	-	-	-	-	N/A
Papakostas et al. [265]	2019	10	-	-	-	-	-	-	x
Yang and Ren [357]	2019	20	-	-	-	-	-	-	-
Wu et al. [353]	2018	N/A	-	-	-	-	-	-	N/A
Baghdadi et al. [26]	2018	20	-	-	-	-	-	-	-
Beéck et al. [33]	2018	29	-	-	-	-	-	-	x
Gordienko et al. [135]	2018	N/A	-	-	-	-	-	-	N/A
Jamaluddin et al. [169]	2018	20	-	-	-	-	-	-	N/A
Karthick et al. [187]	2018	52	-	-	-	-	-	-	N/A
Aryal et al. [22]	2017	12	-	-	-	-	-	-	-
Lopez et al. [237]	2017	19	-	-	-	-	-	-	-
Shahmoradi et al. [302]	2017	6	-	-	-	-	-	-	N/A
Vandewiele et al. [336]	2017	45	-	-	-	-	-	-	x
Buckley et al. [55]	2017	21	-	-	-	-	-	-	x
Maman et al. [243]	2017	8	-	-	-	-	-	x	-
Carey et al. [65]	2016	45	-	-	-	-	-	-	-
Kupschick et al. [211]	2016	22	-	-	-	-	-	-	-
Pernek et al. [274]	2015	11	-	-	-	-	x	-	x
Bilgin et al. [41]	2015	31	-	-	-	-	-	-	-
Karg et al. [186]	2014	7	-	-	-	-	-	-	x
Zhang et al. [367]	2013	17	-	-	-	-	-	-	-
Janssen et al. [170]	2011	9	-	-	-	-	-	-	-
Subasi and Kiymik [319]	2009	14	-	-	-	-	-	-	-
Karg et al. [185]	2008	14	-	-	-	-	-	-	x
This thesis	2024	48	x	x	x	x	x	x	x

Related Works Details

B

Tab. B.1.: Overview of sensors utilised in the related works.

Authors	Exercise	Sensor	Sampling Rate
Albert and Arnrich [9]	Squats	Kinect, ECG, IMU, kMeter	30 Hz
Gan et al. [126]	Squats	ECG	N/A
Huang et al. [161]	Static	sEMG	N/A
Ma and Guo [240]	Yoga	N/A	N/A
Mu et al. [256]	Running	ECG	N/A
Wang et al. [342]	Bicep Curls	sEMG, Fabric sensor, Goniometer	2 kHz, 1 kHz, 1 kHz
Wang et al. [346]	Material Handling	EEG	500 Hz
Yao [358]	N/A	Muscle tone signal	500 Hz
Zhang et al. [365]	Running	IMU, Plantar force	110 Hz
Adapa et al. [1]	Bicep Curls	sEMG	4 kHz
Antwi-Afari et al. [20]	Material Handling	Plantar pressure, IMU	50 Hz
Anwer et al. [21]	Material Handling	ECG, RESP, ST (EQ02 system)	256 Hz
Biró et al. [45]	Cycling, Running, Football	Radar, IMU, ECG	24 GHz, 100 Hz, 9 Hz
Biró et al. [44]	Running	IMU	30 Hz
Bouteraa et al. [52]	Wrist Torque	sEMG	100 kHz
Cañellas et al. [63]	N/A	Thermal	8.7 Hz
Concha-Pérez et al. [83]	Squeeze/Release (Arm)	sEMG, IMU	500 Hz
Dang et al. [92]	Dynamometer	sEMG	2 kHz
De Vito et al. [97]	Material Handling	sEMG	1 kHz
Dimmick et al. [102]	Running	IMU, Garmin devices	1125 Hz
Feng et al. [116]	Rope-Skipping	ECG	512 Hz
Kathirgamanathan et al. [191]	Running	IMU	256 Hz
Liu et al. [235]	Elbow	sEMG	1 kHz
Marena et al. [246]	Material Handling	ECG, Spiroergometria	N/A
Perpetuini et al. [275]	Squats	sEMG, Thermal	250 Hz, 10 Hz
Pirscoveanu and Oliveira [278]	Running	ECG (Smartwatch)	1 Hz
Pravin et al. [280]	Bicep Curls	sEMG	2 kHz
Smiley et al. [313]	Cycling	ECG, PPG, RESP	1024 Hz
Valla et al. [335]	Archimedean Spiral Test	IMU (Smartphone)	N/A
Albert et al. [10]	Squats	IMU, ECG, Azure Kinect, kMeter	128 Hz, 1 kHz, 30 Hz
Bustos et al. [58]	Running	ECG, ST, RESP	N/A
Jaiswal et al. [168]	Walking	ECG, EDA, EMG	N/A
Umer et al. [332]	Material Handling	ECG	N/A
Cheah et al. [69]	Sit-Ups	sEMG, (IMU)	2 kHz
Escobar-Linero et al. [112]	Material Handling	IMUs	51.2 Hz
Guo et al. [139]	Bicep Curls	Blood, ECG	N/A
Jiang et al. [176]	Squats	IMUs	N/A
Li and Chen [225]	Pilates	ECG, sEMG	2 kHz, 2 kHz
Shi et al. [306]	Walking	Angle, ECG, VO2	100 Hz
Triantafyllopoulos et al. [327]	Running	Audio	16 kHz
Wang et al. [343]	Running	IMUs	200 Hz
Zhu et al. [374]	Walking, Cycling, Running	ECG	8 kHz
Chen et al. [71]	Material Handling	ECG, sEMG, PPG	2 kHz
Chen et al. [70]	Material Handling	sEMG, ECG	2 kHz

Table B.1 continued from the previous page

Authors	Exercise	Sensor	Sampling Rate
K et al. [181]	Bicep Curls	sEMG	10 kHz
Sadat-Mohammadi et al. [288]	Material Handling	RESP, IMU	1 kHz
Wang et al. [344]	Cycling	sEMG, RESP	2 kHz
Zhang et al. [370]	Shoulder	sEMG	4 kHz, 1962 Hz
Aguirre et al. [4]	Sit-to-Stand	Kinect, ECG	30 Hz, 1 Hz
Balaskas and Siozios [28]	Running	IMUs	200 Hz
Chalitsios et al. [67]	Running	IMUs, FP, RESP	218 Hz, 516 Hz
Chen et al. [72]	Dumbbell (pick-up)	sEMGs	1 kHz
Elshafei et al. [108]	Bicep Curls	IMU, PPG	50 Hz
Guan et al. [138]	Running	IMU, ECG	50 Hz, 500 Hz
Jiang et al. [175]	Squats, Jacks, Touch	IMUs, FP, MoCap	240 Hz, 100 Hz, 100 Hz
Karvekar et al. [189]	Squats, Walking	IMU	100 Hz
Kuschan and Krüger [212]	Material Handling	IMUs	20 Hz
Lambay et al. [216]	Material Handling	IMU, ECG	N/A
Wang and He [341]	Running	Kinect	30 Hz
Davidson et al. [95]	Running	ECG, GPS, VO2 peak	1 Hz
Luo et al. [238]	Daily Activities	EDA, PPG, ST, IMU, Barometer	1 Hz
Umer et al. [331]	Material Handling	RESP, ST, ECG	25.6 Hz, 0.25 Hz
Wang et al. [345]	Cycling	sEMG, RESP	2 kHz
Guaitolini et al. [137]	Walking, Running	IMUs, MoCap	100 Hz
Maman et al. [242]	Material Handling	IMUs, ECG	25 Hz, 1 kHz
Nasirzadeh et al. [261]	Material Handling	ECG	25 Hz, 50 Hz, 125 Hz
Sani et al. [293]	Material Handling	ECG	N/A
Zhang and Wang [364]	Ball Sports	Camera (eye lid)	N/A
Chowdhury et al. [77]	Walking, Running	ECG , EDA, ST	1 Hz, 4 Hz, 4 Hz
Geurkink et al. [129]	Football	ECG, GPS, IMU	20 Hz, 10 Hz, 100 Hz
Jebelli et al. [171]	Material Handling	PPG, EDA, ST	64 Hz, 4 Hz, 4 Hz
Karvekar et al. [188]	Squats, Walking	IMU	100 Hz
Papakostas et al. [265]	Shoulder	EMG	1926 Hz
Yang and Ren [357]	Muscle Chair	EEG	128 Hz
Wu et al. [353]	Running, Walking, Pedalling	ECG	0.5 Hz
Baghdadi et al. [26]	Material Handling	IMU	51.2 Hz
Beéck et al. [33]	Running	IMUs, ECG	1024 Hz, 1 Hz
Gordienko et al. [135]	Walking, Running, Skiing	IMU, ECG, EEG, GPS	N/A
Jamaluddin et al. [169]	Running	sEMG, ECG	1000 Hz
Karthick et al. [187]	Bicep Curls	sEMG	10 kHz
Aryal et al. [22]	Material Handling	ST, ECG, EEG	N/A
Lopez et al. [237]	Running (stairs)	Thermal	8.7 Hz
Shahmoradi et al. [302]	Reaching (arm)	sEMG, (Kinect)	1 kHz, (30 Hz)
Vandewiele et al. [336]	Football	IMU, GPS, ECG	N/A
Buckley et al. [55]	Running	IMU	256 Hz
Maman et al. [243]	Material Handling	IMUs, ECG	51.2 Hz
Carey et al. [65]	Football	IMU, GPS, ECG	100 Hz, 10 Hz
Kupschick et al. [211]	Material Handling	ECG, ST	N/A
Pernek et al. [274]	Dumbbell (upper body)	IMU	30 Hz
Bilgin et al. [41]	Running	sEMG, MMG, ACC	254–313 Hz, 173–234 Hz
Karg et al. [186]	Squats	MoCap	100 Hz
Zhang et al. [367]	Squats, Walking	IMUs, FP, MoCap	120 Hz, 1200 Hz
Janssen et al. [170]	Leg, Walking	FP, Light barrier	1 kHz
Subasi and Kiymik [319]	Dumbbell	sEMG	1 kHz
Karg et al. [185]	Rowing, Walking	MoCap	240 Hz
This thesis	Squats	IMU, PE	200 Hz, 30 Hz

Tab. B.2.: Overview of the ground truth used in the related works.

Authors	Exercise	Ground Truth	GT Frequency (every ...)
Albert and Arnrich [9]	Squats	RPE20	12 reps
Gan et al. [126]	Squats	RPE10	30/15 reps
Huang et al. [161]	Static	RPE10	N/A
Ma and Guo [240]	Yoga	Blood samples	N/A
Mu et al. [256]	Running	Visual Analog Scale	Session
Wang et al. [342]	Bicep Curls	K-means clustering	-
Wang et al. [346]	Material Handling	Questionnaire	Session (approx. 3 min)
Yao [358]	N/A	K-means clustering	-
Zhang et al. [365]	Running	First vs last 5 min	-
Adapa et al. [1]	Bicep Curls	Activity Intensity	Activity
Antwi-Afari et al. [20]	Material Handling	RPE20	2 min
Anwer et al. [21]	Material Handling	RPE20	15 min
Biró et al. [45]	Cycling, Running, Football	RPE20, Heart Rate	N/A
Biró et al. [44]	Running	Activity Intensity (Beep test)	-
Bouteraa et al. [52]	Wrist Torque	Uncertainty algorithm	-
Cañellas et al. [63]	N/A	Linearly annotated	Session
Concha-Pérez et al. [83]	Squeeze/Release (Arm)	Activity Intensity	-
Dang et al. [92]	Dynamometer	Activity Intensity	-
De Vito et al. [97]	Material Handling	N/A	N/A
Dimmick et al. [102]	Running	RPE, MLSS, first and last km	5 min
Feng et al. [116]	Rope-Skipping	Activity Intensity	-
Kathirgamanathan et al. [191]	Running	Activity Intensity (Beep test)	-
Liu et al. [235]	Elbow	RPE20	N/A
Marena et al. [246]	Material Handling	Metabolic rate	-
Perpetuini et al. [275]	Squats	Activity Intensity	-
Pirscoveanu and Oliveira [278]	Running	RPE20	400 m
Pravin et al. [280]	Bicep Curls	Activity Intensity	N/A
Smiley et al. [313]	Cycling	RPE10	1 min
Valla et al. [335]	Archimedean Spiral Test	Questionnaire	N/A
Albert et al. [10]	Squats	RPE20, lactate	12 reps
Bustos et al. [58]	Running	RPE20	4 min
Jaiswal et al. [168]	Walking	First sets vs last two sets	-
Umer et al. [332]	Material Handling	RPE20	5 min
Cheah et al. [69]	Sit-Ups	First vs last 20% reps	-
Escobar-Linero et al. [112]	Material Handling	RPE20	10 min
Guo et al. [139]	Bicep Curls	RPE	N/A
Jiang et al. [176]	Squats	RPE10	5 reps
Li and Chen [225]	Pilates	RPE	30 s
Shi et al. [306]	Walking	Activity Intensity	-
Triantafyllopoulos et al. [327]	Running	RPE20	3–5 min
Wang et al. [343]	Running	RPE20	100/400 m
Zhu et al. [374]	Walking, Cycling, Running	Activity Intensity	-
Chen et al. [71]	Material Handling	Control group	Session
Chen et al. [70]	Material Handling	Control group	-
K et al. [181]	Bicep Curls	First vs last rep	-
Sadat-Mohammadi et al. [288]	Material Handling	Activity Intensity, NASA-TLX	Activity
Wang et al. [344]	Cycling	Ventilation Threshold	Session
Zhang et al. [370]	Shoulder	10 s after exhaustion	Set
Aguirre et al. [4]	Sit-to-Stand	RPE10	30 s
Balaskas and Siozios [28]	Running	Clustering	-
Chalitsios et al. [67]	Running	Ventilatory Threshold	-

Table B.2 continued from previous page

Authors	Exercise	Ground Truth	GT Frequency (every ...)
Chen et al. [72]	Dumbbell (pick-up)	Manually labelled	N/A
Elshafei et al. [108]	Bicep Curls	RPE20	15 reps
Guan et al. [138]	Running	RPE20	N/A
Jiang et al. [175]	Squats, Jacks, Touch	RPE10	5 reps
Karvekar et al. [189]	Squats, Walking	RPE20	2 min
Kuschan and Krüger [212]	Material Handling	RPE10	Set
Lambay et al. [216]	Material Handling	RPE	10 min
Wang and He [341]	Running	RPE20	2 min
Davidson et al. [95]	Running	RPE20	400/1000 m
Luo et al. [238]	Daily Life	Fatigue Assessment Scale	24 h
Umer et al. [331]	Material Handling	RPE20, SWAT	5 min
Wang et al. [345]	Cycling	Ventilation Threshold	Session
Guaitolini et al. [137]	Walking, Running	First vs other reps	-
Maman et al. [242]	Material Handling	RPE	10 min
Nasirzadeh et al. [261]	Material Handling	RPE20	60 min
Sani et al. [293]	Material Handling	RPE	10 min
Zhang and Wang [364]	Ball Sports	PERCLOS P80	-
Chowdhury et al. [77]	Walking, Running	RPE20	Activity
Geurkink et al. [129]	Football	RPE10	Session
Jebelli et al. [171]	Material Handling	Activity Intensity	-
Karvekar et al. [188]	Squats, Walking	RPE20	2 min
Papakostas et al. [265]	Shoulder	Exhaustion plus 10 s	Set
Yang and Ren [357]	Muscle Chair	RPE10	Set
Wu et al. [353]	Running, Walking, Pedalling	Exercise intensity	-
Baghdadi et al. [26]	Material Handling	RPE20	1/10 min
Beéck et al. [33]	Running	RPE20	400 m
Gordienko et al. [135]	Walking, Running, Skiing	Clustering	-
Jamaluddin et al. [169]	Running	Questionnaire	24 h
Karthick et al. [187]	Bicep Curls	First segments vs last segment	-
Aryal et al. [22]	Material Handling	RPE20	10 m
Lopez et al. [237]	Running (stairs)	Activity Intensity	-
Shahmoradi et al. [302]	Reaching (arm)	Maximum Voluntary Contractions	Set
Vandewiele et al. [336]	Football	RPE10	Session
Buckley et al. [55]	Running	Last 400 m	-
Maman et al. [243]	Material Handling	RPE20	10 min
Carey et al. [65]	Football	RPE10	Session
Kupschick et al. [211]	Material Handling	RPE20	5 min
Pernek et al. [274]	Dumbbell (upper body)	RPE20	10 reps
Bilgin et al. [41]	Running	Bruce protocol	-
Karg et al. [186]	Squats	Questionnaire	5 reps
Zhang et al. [367]	Squats, Walking	Until 60% maximal exertion	Set
Janssen et al. [170]	Leg, Walking	Activity Intensity	-
Subasi and Kiymik [319]	Dumbbell	N/A	N/A
Karg et al. [185]	Rowing, Walking	Activity Intensity	-
This thesis	Squats	RPE20	10 s

Tab. B.3.: Overview of the applied ML models in the related works.

Authors	SVM	RF	ANN	LR	*k*-NN	NB	CNN	LSTM	DT	LDA	RNN	Other	# Models
Albert and Arnrich [9]	-	-	-	-	-	-	-	-	-	-	-	x	4
Gan et al. [126]	x	-	-	-	-	-	-	-	-	-	-	-	1
Huang et al. [161]	x	x	-	-	x	-	x	-	-	-	-	-	4
Ma and Guo [240]	-	-	x	-	-	-	-	-	-	-	-	x	1
Mu et al. [256]	x	-	-	x	x	x	x	x	-	-	x	x	4
Wang et al. [342]	x	-	x	x	x	-	-	-	-	-	-	-	4
Wang et al. [346]	-	-	-	-	-	-	x	-	-	-	-	x	9
Yao [358]	x	x	-	x	x	-	-	-	-	-	-	-	5
Zhang et al. [365]	x	-	-	-	x	x	-	-	x	-	-	-	4
Adapa et al. [1]	x	x	-	x	-	-	-	-	-	-	-	x	5
Antwi-Afari et al. [20]	x	x	x	-	x	-	-	-	x	-	-	-	5
Anwer et al. [21]	x	x	x	-	x	-	-	-	x	-	-	-	5
Biró et al. [45]	-	-	-	x	-	-	-	-	-	-	-	-	1
Biró et al. [44]	x	x	-	x	x	x	-	x	x	x	-	x	14
Bouteraa et al. [52]	x	-	-	-	-	-	-	-	-	-	-	-	1
Cañellas et al. [63]	-	-	-	-	-	-	x	-	-	-	-	x	7
Concha-Pérez et al. [83]	-	-	-	-	-	-	-	-	-	-	-	x	1
Dang et al. [92]	-	-	-	-	x	-	-	x	-	-	-	x	4
De Vito et al. [97]	-	-	-	-	-	-	-	-	x	-	-	x	1
Dimmick et al. [102]	-	x	-	-	-	-	-	-	-	-	-	-	1
Feng et al. [116]	-	x	-	-	x	-	-	-	-	-	-	x	6
Kathirgamanathan et al. [191]	-	-	-	-	x	-	-	-	-	-	-	x	2
Liu et al. [235]	-	x	-	-	-	-	x	x	-	-	-	x	4
Marena et al. [246]	x	-	x	x	-	-	-	-	x	-	-	x	6
Perpetuini et al. [275]	-	-	-	x	-	-	-	-	-	-	-	x	5
Pirscoveanu and Oliveira [278]	x	-	x	x	-	-	-	-	x	-	-	x	6
Pravin et al. [280]	x	x	-	x	-	-	-	-	-	-	-	-	3
Smiley et al. [313]	x	-	x	-	-	x	-	-	x	-	-	x	5
Valla et al. [335]	x	x	-	x	x	-	-	-	x	-	-	x	6
Albert et al. [10]	-	x	-	-	-	-	-	-	-	-	-	x	4
Bustos et al. [58]	x	x	x	-	x	-	-	-	-	-	-	x	6
Jaiswal et al. [168]	x	x	-	x	-	-	-	x	-	-	x	-	4
Umer et al. [332]	-	-	x	-	-	-	-	-	x	-	-	x	6
Cheah et al. [69]	-	-	-	-	-	-	-	-	-	-	-	x	1
Escobar-Linero et al. [112]	x	-	x	-	-	-	-	x	-	-	x	-	4
Guo et al. [139]	x	-	x	-	-	-	-	-	x	-	-	-	3
Jiang et al. [176]	-	-	-	-	-	-	x	x	-	-	-	x	3
Li and Chen [225]	x	-	x	-	x	-	-	-	-	x	-	-	4
Shi et al. [306]	x	x	-	x	-	-	-	-	x	-	-	x	5
Triantafyllopoulos et al. [327]	-	-	-	-	-	-	x	-	-	-	-	-	1
Wang et al. [343]	x	x	-	-	-	-	-	-	-	-	-	-	2
Zhu et al. [374]	-	-	-	-	-	-	x	-	-	-	-	-	1
Chen et al. [71]	x	x	-	-	-	-	-	-	-	-	-	-	2
Chen et al. [70]	x	x	-	-	-	-	-	-	-	-	-	x	3
K et al. [181]	-	-	-	-	-	-	-	-	-	-	-	x	1
Sadat-Mohammadi et al. [288]	x	x	x	-	x	-	-	-	-	-	-	-	4
Wang et al. [344]	x	-	-	-	-	-	x	x	-	-	-	-	3
Zhang et al. [370]	x	x	-	-	x	-	x	x	-	-	-	x	1
Aguirre et al. [4]	x	x	x	x	x	-	-	-	-	-	-	-	5
Balaskas and Siozios [28]	x	-	x	-	-	-	-	x	-	-	-	-	3
Chalitsios et al. [67]	-	x	-	-	-	-	-	-	-	-	-	-	1

Table B.3 continued from previous page

Authors	SVM	RF	ANN	LR	*k*-NN	NB	CNN	LSTM	DT	LDA	RNN	Other	# Models
Chen et al. [72]	-	0	-	-	-	-	-	x	-	-	-	-	1
Elshafei et al. [108]	-	x	x	x	-	-	-	-	x	-	-	x	5
Guan et al. [138]	x	x	x	-	-	-	-	x	-	-	-	-	4
Jiang et al. [175]	-	x	-	-	-	-	x	-	-	-	-	-	2
Karvekar et al. [189]	x	-	-	-	-	-	-	-	-	-	-	-	1
Kuschan and Krüger [212]	x	-	-	-	-	-	-	-	-	-	-	-	1
Lambay et al. [216]	-	-	-	-	-	-	-	x	-	-	x	-	2
Wang and He [341]	x	x	-	-	-	-	x	-	-	-	-	x	4
Davidson et al. [95]	x	-	-	-	x	-	x	-	-	-	-	x	5
Luo et al. [238]	-	x	-	-	-	-	x	-	-	-	-	x	2
Umer et al. [331]	x	-	-	-	x	-	-	-	x	-	-	x	5
Wang et al. [345]	x	-	-	-	-	-	x	-	-	-	-	x	4
Guaitolini et al. [137]	x	x	-	-	x	x	-	-	x	-	-	-	5
Maman et al. [242]	x	x	-	x	-	-	-	-	-	-	-	-	3
Nasirzadeh et al. [261]	-	x	x	x	x	x	-	-	x	x	-	x	8
Sani et al. [293]	-	-	-	-	x	-	-	-	-	-	-	-	1
Zhang and Wang [364]	x	-	-	-	-	-	x	-	-	-	-	-	2
Chowdhury et al. [77]	x	x	x	-	-	-	-	-	-	-	-	-	3
Geurkink et al. [129]	-	x	-	x	-	-	-	-	x	-	-	x	6
Jebelli et al. [171]	x	-	-	-	-	-	-	-	-	-	-	-	1
Karvekar et al. [188]	x	-	-	-	-	-	-	-	-	-	-	-	1
Papakostas et al. [265]	x	x	-	-	-	-	-	-	-	-	-	x	3
Yang and Ren [357]	x	-	-	-	-	-	-	-	-	-	-	x	2
Wu et al. [353]	x	-	x	-	x	x	-	-	x	-	-	-	5
Baghdadi et al. [26]	x	-	-	-	-	-	-	-	-	-	-	-	1
Beéck et al. [33]	-	-	x	x	-	-	-	-	-	-	-	x	3
Gordienko et al. [135]	-	-	x	x	-	-	-	-	-	-	-	-	2
Jamaluddin et al. [169]	-	-	-	-	-	x	-	-	-	-	-	-	1
Karthick et al. [187]	x	x	-	-	-	x	-	-	-	-	-	x	4
Aryal et al. [22]	x	-	-	-	-	-	-	-	x	-	-	x	10
Lopez et al. [237]	x	-	-	-	-	-	x	-	-	-	-	x	2
Shahmoradi et al. [302]	-	-	x	-	-	-	-	-	-	-	-	x	2
Vandewiele et al. [336]	-	x	-	x	-	-	-	-	x	-	-	x	5
Buckley et al. [55]	x	x	-	-	x	x	-	-	-	-	-	-	4
Maman et al. [243]	-	-	-	x	-	-	-	-	-	-	-	-	1
Carey et al. [65]	x	x	x	x	-	x	-	-	-	-	-	x	8
Kupschick et al. [211]	x	-	-	-	-	-	-	-	-	-	-	x	2
Pernek et al. [274]	x	-	-	-	-	-	-	-	-	-	-	x	1
Bilgin et al. [41]	-	-	x	-	-	-	-	-	-	-	-	-	1
Karg et al. [186]	-	-	-	x	-	-	-	-	-	-	-	x	2
Zhang et al. [367]	x	-	-	-	-	-	-	-	-	-	-	-	1
Janssen et al. [170]	x	-	-	-	-	-	-	-	-	-	-	x	2
Subasi and Kiymik [319]	-	-	x	-	-	-	-	-	-	-	-	-	1
Karg et al. [185]	x	-	-	-	x	x	-	-	-	x	-	-	4
This thesis	x	x	x	x	x	-	-	-	-	-	-	x	6

Tab. B.4.: Overview of the number of classes, features, and samples as well as what evaluation types and whether cross-evaluation (CV) was applied in the related works.

Authors	Classes	Features	Samples	*T1-SOLO*	*T2-LNSO*	*T3-LOSO*	*T4-LMSO*	CV	Test Ratio (%)
Albert and Arnrich [9]	16	50, 100	2304	-	-	x	-	x	6.3
Gan et al. [126]	3	18	31	N/A	N/A	N/A	N/A	10	N/A
Huang et al. [161]	2	18	N/A	-	x	-	-	N/A	20
Ma and Guo [240]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Mu et al. [256]	2	12	2560	-	x	-	-	5	15
Wang et al. [342]	3	4	580	-	x	-	-	x	30
Wang et al. [346]	4 (5)	73728	483 (12558)	-	x	-	-	N/A	15
Yao [358]	2 (3)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Zhang et al. [365]	2	10	5400	-	-	x	-	5	5.5
Adapa et al. [1]	2	30	N/A	-	-	x	-	N/A	35
Antwi-Afari et al. [20]	2, 3, 4	38	1289	-	x	-	-	10	10
Anwer et al. [21]	4	25	1425	-	x	-	-	10	15
Biró et al. [45]	N/A	N/A	N/A	-	-	x	-	x	N/A
Biró et al. [44]	2	N/A	1201	N/A	N/A	N/A	N/A	N/A	10
Bouteraa et al. [52]	2	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cañellas et al. [63]	101	N/A	418813	N/A	N/A	N/A	N/A	5	N/A
Concha-Pérez et al. [83]	2	126	N/A	-	-	x	x	x	N/A
Dang et al. [92]	3	Raw Data	N/A	-	x	-	-	3	33.3
De Vito et al. [97]	2	22	5634	-	x	-	-	5	10
Dimmick et al. [102]	2	39, 6	N/A	x	-	x	-	x	11.1
Feng et al. [116]	2	28	N/A	-	x	-	-	10	10
Kathirgamanathan et al. [191]	2	1	5510	x	-	x	-	x	33.3, 5.3
Liu et al. [235]	4	2	7560	-	x	-	-	10	30
Marena et al. [246]	N/A	2	N/A	-	-	x	-	5	20
Perpetuini et al. [275]	2	9	N/A	-	-	x	-	x	10
Pirscoveanu and Oliveira [278]	14	about 9	N/A	x	-	x	-	5	12
Pravin et al. [280]	2	6	24	N/A	N/A	N/A	N/A	N/A	N/A
Smiley et al. [313]	2	68	150	-	x	-	-	N/A	20
Valla et al. [335]	2	60	33	-	x	-	-	5	33.3
Albert et al. [10]	14	8	N/A	-	-	x	-	x	8,3
Bustos et al. [58]	4	21	750	-	x	x	-	10	10
Jaiswal et al. [168]	2	169	N/A	-	-	-	x (5)	5	15
Umer et al. [332]	14, 4	>20	1286	-	x	-	-	10	15
Cheah et al. [69]	2	6	1092	-	x	-	-	N/A	20
Escobar-Linero et al. [112]	4	40	360	-	x	-	-	N/A	20
Guo et al. [139]	3	162	800	-	x	-	-	N/A	20
Jiang et al. [176]	10	32	N/A	-	-	x	-	x	1, 6
Li and Chen [225]	3	11	1200	-	x	-	-	x	N/A
Shi et al. [306]	5	12	N/A	-	-	-	-	x	N/A
Triantafyllopoulos et al. [327]	14	5	N/A	-	-	x	-	N/A	21
Wang et al. [343]	3	11	N/A	-	x	-	-	x	10
Zhu et al. [374]	6	N/A	14400	-	x	-	-	N/A	10
Chen et al. [71]	2	16	80	-	x	-	-	5	25
Chen et al. [70]	2	24	94	-	x	-	-	10	10
K et al. [181]	2	7	116	-	x	-	-	10	10
Sadat-Mohammadi et al. [288]	3	10	N/A	-	x	-	-	5	20
Wang et al. [344]	2	4	8872	-	x	-	-	N/A	20–40
Zhang et al. [370]	2	32	18740, 93998	-	x	x	-	x	10
Aguirre et al. [4]	3	33	660	-	-	x	-	6	16, 7

Table B.4 continued from previous page

Authors	Classes	Features	Samples	*T1-SOLO*	*T2-LNSO*	*T3-LOSO*	*T4-LMSO*	CV	Test Ratio (%)
Balaskas and Siozios [28]	2	18	N/A	-	-	-	-	-	N/A
Chalitsios et al. [67]	2	10	29650	-	x	-	-	N/A	30
Chen et al. [72]	2	N/A	5000	-	x	-	-	N/A	30
Elshafei et al. [108]	2	22	3000	x	-	x	-	10	5
Guan et al. [138]	3	88	N/A	-	-	x	-	N/A	7, 2
Jiang et al. [175]	10	10	1790, 1240, 1140	-	-	x	-	5, 6	15
Karvekar et al. [189]	2, 3, 4	42	1240, 1800, 2400	-	x	-	-	5	N/A
Kuschan and Krüger [212]	3, 5	7	282	N/A	N/A	N/A	N/A	5	N/A
Lambay et al. [216]	2	23	N/A	-	-	x	-	x	30
Wang and He [341]	4	117	6624	x	x	x	-	5	25
Davidson et al. [95]	2	3 (Raw Data)	112	-	-	-	x (2)	5, 6	20
Luo et al. [238]	2	254	N/A	-	x	-	-	x	10
Umer et al. [331]	14	7	1286	x	x	-	-	10	10
Wang et al. [345]	2	6	100	-	-	-	x (4)	10	20
Guaitolini et al. [137]	2	6	26	-	-	x	-	x	N/A
Maman et al. [242]	2	7	234 (46800)	-	-	-	x (2)	10	13
Nasirzadeh et al. [261]	2	10	3456, 1728, 691	-	x	-	-	10	10
Sani et al. [293]	2	6	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Zhang and Wang [364]	2	128	8000	-	-	x	-	5	N/A
Chowdhury et al. [77]	3	15	615	-	-	x	-	x	4.6
Geurkink et al. [129]	10	70	913	-	x	-	-	5	20.6
Jebelli et al. [171]	2, 3	57	N/A	-	-	x	-	10	10
Karvekar et al. [188]	2, 4	12	N/A	N/A	N/A	N/A	N/A	-	N/A
Papakostas et al. [265]	2	26	90	x	-	x	-	N/A	N/A
Yang and Ren [357]	2	Variable	220	-	x	-	-	5, 10	20
Wu et al. [353]	2	>5	148	N/A	N/A	N/A	N/A	N/A	N/A
Baghdadi et al. [26]	2	2	1000	-	x	-	-	5	20
Beéck et al. [33]	14	15	7607	x	x	x	-	N/A	N/A
Gordienko et al. [135]	N/A	5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Jamaluddin et al. [169]	2	4	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Karthick et al. [187]	2	12	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Aryal et al. [22]	4	21	253	-	x	-	-	10	10
Lopez et al. [237]	2	4096	5700	-	-	x	-	x	5.3
Shahmoradi et al. [302]	3	8	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Vandewiele et al. [336]	10	28	913	-	x	x	-	5	N/A
Buckley et al. [55]	2	15	584	x	-	x	-	10	N/A
Maman et al. [243]	2	16	144	-	-	-	x (2)	N/A	N/A
Carey et al. [65]	10, 15	23	3398	-	x	-	-	5	N/A
Kupschick et al. [211]	2	4, 6	533	-	-	x	-	x	N/A
Pernek et al. [274]	14	7	264	-	x	x	-	10	25, 9.1
Bilgin et al. [41]	2	2	N/A	-	x	-	-	N/A	35.5
Karg et al. [186]	5	17	445	-	x	x	-	x	N/A
Zhang et al. [367]	2	11	340	-	x	-	-	5	30
Janssen et al. [170]	2, 3	6	162	-	x	-	-	x	N/A
Subasi and Kiymik [319]	2	150–175	1100	-	x	-	-	N/A	45
Karg et al. [185]	2	531	N/A	-	x	x	-	N/A	N/A
This thesis	3	212	3595	-	x	x	x (2–7)	5	2–15

Note: In column *T4-LMSO*, the number in parenthesis represents the number of subjects that were left out for testing.

Tab. B.5.: Mean results (rounded) of the best performing of ML model in the related works.

Authors	Acc (%)	Spec (%)	Recall (%)	Prec (%)	F_1 (%)	CM	RMSE	MAE	MAPE	R^2	Other
Albert and Arnrich [9]	-	-	-	-	-	-	1.5	1.3	8.1	0.2	x
Gan et al. [126]	80.7	-	77.6	76.2	71.9 [sic!]	-	-	-	-	-	x
Huang et al. [161]	86.1	-	-	-	84.6	-	-	-	-	-	-
Ma and Guo [240]	N/A	N/A	N/A	N/A	N/A	-	N/A	N/A	N/A	N/A	x
Mu et al. [256]	94	-	89.3	99	93.7	x	-	-	-	-	x
Wang et al. [342]	83.3	-	90	-	95	x	-	-	-	-	x
Wang et al. [346]	94.7	-	94.4	95	94.7	x	-	-	-	-	x
Yao [358]	83.1	-	89.8	-	94.3	-	-	-	-	-	x
Zhang et al. [365]	99	-	-	-	-	-	-	-	-	-	-
Adapa et al. [1]	86.5	-	-	-	-	-	-	-	-	-	-
Antwi-Afari et al. [20]	96.9	94.73	98.6	96	97.3	x	-	-	-	-	-
Anwer et al. [21]	93.5	-	93.5	93.6	93.5	x	-	-	-	-	x
Biró et al. [45]	>90	-	-	-	-	-	-	-	-	-	x
Biró et al. [44]	59	-	59	59	59	x	-	-	-	-	x
Bouteraa et al. [52]	92.6	89.5	82.1	-	-	-	-	-	-	-	-
Cañellas et al. [63]	-	-	-	-	-	-	13.6	16.5	-	-	x
Concha-Pérez et al. [83]	95.7	-	95.7	91.1	93.3	x	-	-	-	-	-
Dang et al. [92]	93.5	-	44.3	98.8	64.3	-	-	-	-	-	-
De Vito et al. [97]	83.9	-	-	-	77.7	x	-	-	-	-	-
Dimmick et al. [102]	68.9	-	-	-	-	-	-	-	-	-	x
Feng et al. [116]	-	-	-	-	-	-	0.3	-	11.5	0.89	x
Kathirgamanathan et al. [191]	97, 59	-	-	-	-	x	-	-	-	-	-
Liu et al. [235]	96.7	-	96.7	96.7	96.7	x	-	-	-	-	-
Marena et al. [246]	-	-	-	-	-	-	0.5	0.6	-	0.9	-
Perpetuini et al. [275]	-	-	-	-	-	-	0.5	-	-	-	x
Pirscoveanu and Oliveira [278]	-	-	-	-	-	-	0.5, 1.8	-	-	-	x
Pravin et al. [280]	87.5	92	81.8	-	-	x	-	-	-	-	-
Smiley et al. [313]	80	-	-	-	79	-	-	-	-	-	-
Valla et al. [335]	78.8	25	96	80	87.3	x	-	-	-	-	-
Albert et al. [10]	-	-	-	-	-	x	2.2	-	7.7	0.5	x
Bustos et al. [58]	88	-	88	89	88	x	-	-	-	-	-
Jaiswal et al. [168]	80.5	-	88	-	-	-	-	-	-	-	-
Umer et al. [332]	64.2, 75.7	-	-	-	-	x	1.7	-	-	-	x
Cheah et al. [69]	65.3	-	-	-	-	-	-	-	-	-	-
Escobar-Linero et al. [112]	91	-	-	-	-	x	-	-	-	-	-
Guo et al. [139]	92	-	-	-	-	x	-	-	-	-	-
Jiang et al. [176]	83.7	-	-	-	82	-	-	-	-	-	-
Li and Chen [225]	94.3	-	-	-	-	-	-	-	-	-	-
Shi et al. [306]	88.9	-	-	-	-	-	-	-	-	-	-
Triantafyllopoulos et al. [327]	-	-	-	-	-	-	-	2.4	-	-	-
Wang et al. [343]	91.1	-	-	-	-	x	-	-	-	-	-
Zhu et al. [374]	97.7	-	-	-	-	x	-	-	-	-	-
Chen et al. [71]	72	90.9	88.9	88.9	88.9	-	-	-	-	-	x
Chen et al. [70]	89.47	100	100	80	88.9	-	-	-	-	-	x
K et al. [181]	94	95	93.3	-	93.8	-	-	-	-	-	x
Sadat-Mohammadi et al. [288]	93.4	-	-	-	-	x	-	-	-	-	x
Wang et al. [344]	95.2	96.4	94.1	96.7	-	x	-	-	-	-	-
Zhang et al. [370]	96.5, 78.3	-	-	96.4, 78.6	96.5, 77	-	-	-	-	-	-
Aguirre et al. [4]	83.2	-	-	-	82.7	-	-	-	-	-	-
Balaskas and Siozios [28]	43	-	-	-	-	-	-	-	-	-	-
Chalitsios et al. [67]	91.4	89	93	-	-	-	-	-	-	-	x

Table B.5 continued from the previous page

Authors	Acc (%)	Spec (%)	Recall (%)	Prec (%)	F_1 (%)	CM	RMSE	MAE	MAPE	R^2	Other
Chen et al. [72]	90.4	-	-	-	-	-	-	-	-	-	-
Elshafei et al. [108]	18–95	-	-	-	34–87	-	-	-	-	-	-
Guan et al. [138]	80.6	-	-	-	76.9	-	-	-	-	-	-
Jiang et al. [175]	-	-	-	-	-	-	-	-	-	-	x
Karvekar et al. [189]	91, 78, 64	-	-	-	-	x	-	-	-	-	-
Kuschan and Krüger [212]	83.8, 80.9	-	-	-	-	x	-	-	-	-	-
Lambay et al. [216]	65	-	-	-	-	x	-	-	-	-	-
Wang and He [341]	87.7	-	-	-	-	-	-	-	-	-	-
Davidson et al. [95]	84.8	-	84.8	85.1	84.9	x	-	-	-	-	x
Luo et al. [238]	71.4	-	73	70	71	-	-	-	-	-	x
Umer et al. [331]	98.5, 95.3	-	-	-	-	x	-	-	-	-	x
Wang et al. [345]	83.5	-	-	-	-	-	-	-	-	-	-
Guaitolini et al. [137]	84.6	-	-	-	-	x	-	-	-	-	-
Maman et al. [242]	>85	-	-	-	-	x	-	-	-	-	-
Nasirzadeh et al. [261]	90.36	-	-	-	-	-	-	-	-	-	-
Sani et al. [293]	78.2	-	-	-	-	-	-	-	-	-	-
Zhang and Wang [364]	90	-	-	-	-	-	-	-	-	-	-
Chowdhury et al. [77]	-	-	-	-	85.2	x	-	-	-	-	-
Geurkink et al. [129]	91.7	-	-	-	-	x	0.9	0.7	-	-	-
Jebelli et al. [171]	90, 87	-	-	-	-	x	-	-	-	-	x
Karvekar et al. [188]	91, 61	-	-	-	-	x	-	-	-	-	-
Papakostas et al. [265]	-	-	-	-	77.8, 70.4	-	-	-	-	-	-
Yang and Ren [357]	90	-	-	-	-	-	-	-	-	-	-
Wu et al. [353]	98.7	-	-	-	-	-	-	-	-	-	-
Baghdadi et al. [26]	90	-	-	-	-	-	-	-	-	-	-
Beéck et al. [33]	-	-	-	-	-	-	-	1.7–1.9	-	-	-
Gordienko et al. [135]	-	-	-	-	-	-	-	0.2	-	-	-
Jamaluddin et al. [169]	98	100	96	-	-	-	-	-	-	-	-
Karthick et al. [187]	91.5	93.9	89.1	-	-	-	-	-	-	-	-
Aryal et al. [22]	82.6	-	-	-	-	x	-	-	-	-	-
Lopez et al. [237]	81.5	-	-	-	-	-	-	-	-	-	x
Shahmoradi et al. [302]	95.3	-	-	-	-	x	-	-	-	-	-
Vandewiele et al. [336]	-	-	-	-	-	-	0.7, 0.9	-	-	-	x
Buckley et al. [55]	75	77	73	-	74.9	-	-	-	-	-	-
Maman et al. [243]	-	88, 89	96, 95	-	-	-	-	-	-	-	-
Carey et al. [65]	-	-	-	-	-	-	1.0	-	-	-	-
Kupschick et al. [211]	85.8	-	-	-	-	-	-	-	-	-	-
Pernek et al. [274]	-	-	-	-	-	-	-	-	-	-	x
Bilgin et al. [41]	92	-	-	-	-	-	-	-	-	-	-
Karg et al. [186]	81	-	-	-	-	-	0.6	-	-	-	-
Zhang et al. [367]	90	-	-	-	-	-	-	-	-	-	-
Janssen et al. [170]	98.1	-	-	-	-	-	-	-	-	-	-
Subasi and Kiymik [319]	91	87	90	-	-	-	-	-	-	-	x
Karg et al. [185]	100	-	-	-	-	-	-	-	-	-	-
This thesis	78.4	80.4	76.3	-	78.3	x	1.1	2.6	-	-	x

Note: CM stands for confusion matrix.

Tab. B.6.: Overview of the balance of classes in the related works.

Authors	Folds stratified	Classes Balanced	Oversampled	Undersampled	Notes
Albert and Arnrich [9]	-	Yes	x	-	-
Gan et al. [126]	-	No	-	-	-
Huang et al. [161]	-	N/A	N/A	N/A	-
Ma and Guo [240]	-	N/A	N/A	N/A	-
Mu et al. [256]	-	N/A	-	-	-
Wang et al. [342]	-	No	-	-	-
Wang et al. [346]	-	Nearly	x	x	The number of samples for level 5 was lower due to individual differences
Yao [358]	-	N/A	-	-	-
Zhang et al. [365]	-	Yes	-	-	-
Adapa et al. [1]	-	No	-	25%	Different duration of reps
Antwi-Afari et al. [20]	-	Nearly	-	-	-
Anwer et al. [21]	-	N/A	-	-	-
Biró et al. [45]	-	N/A	x	-	-
Biró et al. [44]	-	Yes	-	-	-
Bouteraa et al. [52]	-	N/A	-	-	-
Cañellas et al. [63]	-	Yes	x	-	-
Concha-Pérez et al. [83]	-	N/A	-	-	-
Dang et al. [92]	-	Yes	-	-	Fatigue class is in majority
De Vito et al. [97]	-	No	-	x	-
Dimmick et al. [102]	-	No, Yes	-	-	Experiment 1 and 2
Feng et al. [116]	-	N/A	-	-	-
Kathirgamanathan et al. [191]	-	Yes	-	-	-
Liu et al. [235]	-	Yes	-	-	-
Marena et al. [246]	-	N/A	-	-	-
Perpetuini et al. [275]	-	N/A	-	-	-
Pirscoveanu and Oliveira [278]	-	N/A	-	-	-
Pravin et al. [280]	-	N/A	-	-	-
Smiley et al. [313]	-	N/A	-	-	-
Valla et al. [335]	-	No	-	-	-
Albert et al. [10]	-	No	-	-	-
Bustos et al. [58]	x	No	-	-	-
Jaiswal et al. [168]	x	No	-	-	-
Umer et al. [332]	-	No	-	-	-
Cheah et al. [69]	-	N/A	-	-	-
Escobar-Linero et al. [112]	-	No	-	-	RPE deltas used to reduce and balance classes but still fewer samples for edge classes
Guo et al. [139]	-	N/A	-	-	-
Jiang et al. [176]	-	N/A	-	-	-
Li and Chen [225]	-	No	-	-	-
Shi et al. [306]	-	N/A	-	-	-
Triantafyllopoulos et al. [327]	-	No	-	-	Fewer samples for low and high fatigue classes
Wang et al. [343]	-	N/A	-	-	-
Zhu et al. [374]	-	Yes	-	x	Poor quality data discarded
Chen et al. [71]	-	N/A	-	-	-
Chen et al. [70]	-	N/A	-	-	"Skewed values"
K et al. [181]	-	Yes	-	-	-
Sadat-Mohammadi et al. [288]	-	N/A	-	-	-
Wang et al. [344]	-	Nearly	-	-	-
Zhang et al. [370]	-	N/A	-	-	-

Table B.6 continued from the previous page

Authors	Folds stratified	Classes Balanced	Oversampled	Undersampled	Notes
Aguirre et al. [4]	-	Yes	-	-	Different number of samples per subject
Balaskas and Siozios [28]	-	N/A	-	-	-
Chalitsios et al. [67]	-	Yes	-	-	Backward data selection
Chen et al. [72]	-	N/A	-	-	-
Elshafei et al. [108]	-	No	-	-	Later sets have a longer duration
Guan et al. [138]	-	Yes	SMOTE	-	Fatigue is the minority class
Jiang et al. [175]	-	Yes	Duplicates	-	Number of sets vary between subjects
Karvekar et al. [189]	-	N/A	-	-	-
Kuschan and Krüger [212]	-	No	-	-	Fatigue class in minority
Lambay et al. [216]	-	N/A	-	-	-
Wang and He [341]	-	No	-	-	Large gap between fatigue classes proportions. One subject was acquired 4 times
Davidson et al. [95]	-	N/A	x	-	-
Luo et al. [238]	-	No	-	-	-
Umer et al. [331]	-	No	-	-	-
Wang et al. [345]	-	N/A	-	-	-
Guaitolini et al. [137]	-	N/A	-	-	-
Maman et al. [242]	-	Yes	-	20%	Subjects varied in age and experience.
Nasirzadeh et al. [261]	-	Yes	-	-	-
Sani et al. [293]	-	N/A	-	-	-
Zhang and Wang [364]	-	Yes	-	66.7%	-
Chowdhury et al. [77]	-	No	-	-	-
Geurkink et al. [129]	-	No	-	-	-
Jebelli et al. [171]	-	N/A	-	-	-
Karvekar et al. [188]	-	N/A	-	-	-
Papakostas et al. [265]	-	N/A	-	-	-
Yang and Ren [357]	-	N/A	-	-	-
Wu et al. [353]	-	N/A	-	-	Normal distribution
Baghdadi et al. [26]	-	Yes	-	-	25 strides of each set. Subject included if RPE greater than 10
Beéck et al. [33]	-	No	-	-	Different number of samples per subject
Gordienko et al. [135]	-	N/A	-	-	-
Jamaluddin et al. [169]	-	N/A	-	-	-
Karthick et al. [187]	-	Yes	-	-	First and last curl
Aryal et al. [22]	-	N/A	-	-	-
Lopez et al. [237]	-	N/A	-	-	-
Shahmoradi et al. [302]	-	N/A	-	-	-
Vandewiele et al. [336]	-	No	-	-	Normal distribution
Buckley et al. [55]	-	Yes	-	-	-
Maman et al. [243]	-	Yes	SMOTE	x	Accuracy is inappropriate due to class imbalances. Random Under Sampling
Carey et al. [65]	-	N/A	-	-	-
Kupschick et al. [211]	-	N/A	-	-	-
Pernek et al. [274]	x	N/A	-	-	-
Bilgin et al. [41]	-	N/A	-	-	-
Karg et al. [186]	-	Yes	-	x	Fewer samples for low and high fatigue classes
Zhang et al. [367]	-	Yes	-	-	-
Janssen et al. [170]	-	No	-	-	Different number of samples per subject
Subasi and Kiymik [319]	-	Yes	-	-	-
Karg et al. [185]	-	N/A	-	-	-
This thesis	-	Yes	SMOTE	-	Fatigue class is minority. More male subjects (2:1)

Tab. B.7.: Generalisation in the related works.

Authors	Generalisation / Variance / Small Data / Limitations
Albert and Arnrich [9]	Large window overlap of 95%. Small dataset limits the generalisation of the results. Weak outcomes for 5 of 16 subjects. RPE labels grouped for oversampling and to balance the distribution.
Gan et al. [126]	Larger sample population to assess stability.
Huang et al. [161]	N/A
Ma and Guo [240]	Expand sample size for different groups and environmental conditions to enhance generalisability. Feature selection improve performance and generalisability. Splitting data into training and test set prevent overfitting problems and improves generalisability.
Mu et al. [256]	Develop hierarchical transformer for learning small time series
Wang et al. [342]	Vague boundary between non-fatigue and fatigue. SVM can be effective for small datasets with strong generalisation ability. Cross-validation to access how well the model generalise.
Wang et al. [346]	Less observations for level 5 due to individual differences among participants. Level 5 fatigue was removed to improve generalisability. Due to small dataset, augmentation was applied to enhance the model's ability to generalise to new data.
Yao [358]	Universality of the model needs to be strengthened. More comprehensive data needed to ensure the adequacy and representativeness of the training set.
Zhang et al. [365]	SVM is the most widely used model due to its suitability for small samples. Generalisability accessed by using *T3-LOSO*.
Adapa et al. [1]	N/A
Antwi-Afari et al. [20]	There are differences in human characteristics between students and construction workers. Huge samples of data would help to generalise the findings.
Anwer et al. [21]	Duration of the experiments was not a typical half-day of construction work. Compared to many strong learners which tend to remember data and overfit, bagging reduces amount of variation in a dataset and reduce the amount of overfitting.
Biró et al. [45]	N/A
Biró et al. [44]	More variability in fatigue state. Fluctuation in stamina values across strides varies between participants. Individual profiles mean that models must be personalised being data-intensive and time-consuming. There is a risk of overfitting. LSTM or Gated Recurrent Units would improve generalisability.
Bouteraa et al. [52]	N/A
Cañellas et al. [63]	N/A
Concha-Pérez et al. [83]	N/A
Dang et al. [92]	N/A
De Vito et al. [97]	Bootstrap and bagging helps to reduce overfitting and improve generalisation of the ensemble model.
Dimmick et al. [102]	Random Forest allows for robustness to small datasets. Subject-specific models were more accurate than group-based models. Different feature importance ranking were observed between different subjects. Differences in fitness/experience could explain the high variability in subject-specific model accuracies. Risk of overfitting due to small sample size, although cross-validation was applied to reduce the risk.
Feng et al. [116]	Small sample size is a general limitation due to high cost of hiring subjects.
Kathirgamanathan et al. [191]	Personalised classifiers perform up to 40% better than group-based models. For this reason, similar subjects are clustered into groups. The clustered groups perform up to 20% better than random clusters. *T1-SOLO* achieved 97% and *T3-LOSO* 59%.
Liu et al. [235]	Adjacent data samples of fatigue were prone to confusion during classification. For this reason, neighbouring samples were removed.
Marena et al. [246]	All subjects were college-aged individuals and self-reported fit. To enhance generalisability, broader range of subjects with various age and fitness levels is needed. Activities with greater physical demand are more predictable.
Perpetuini et al. [275]	Larger sample size for more generalisable estimation. Cross-validation was applied to assess generalisability. A larger data set would provide more generalisable estimation.
Pirscoveanu and Oliveira [278]	Low-quality predictions may reach high errors due to extended periods of identical RPE during measurements.
Pravin et al. [280]	N/A
Smiley et al. [313]	Larger sample size for additional validations.
Valla et al. [335]	Low specificity is due to fatigued subjects exhibiting less variation. In future, we intend to develop personalised models. Bias may lead to overfitting which is avoided by cross-validation.
Albert et al. [10]	Small sample size is a limitation. Only male subjects for a homogenous population.
Bustos et al. [58]	Varying accuracy for the four classes from 69 to 93% with *T3-LOSO*. 6% difference between *T2-LNSO* and *T3-LOSO* indicates potential overfitting when randomising the data.
Jaiswal et al. [168]	N/A

Table B.7 continued from the previous page

Authors	Generalisation / Variance / Small Data / Limitations
Umer et al. [332]	Varying perceived physical exertion and responses. Combining multiple levels of exertion lead to better accuracy. Variability of perceived exertion and physiological responses might limit accuracy. Multi-class multi-task models could learn generalised and individualised features for better predictions despite inter-personal variability.
Cheah et al. [69]	Feature space reduction to reduce number of required samples.
Escobar-Linero et al. [112]	N/A
Guo et al. [139]	N/A
Jiang et al. [176]	Subjects are divided into fast-tiring and slow-tiring subgroups. No significant difference was observed.
Li and Chen [225]	N/A
Shi et al. [306]	Deep learning usually requires a large amount of data. SVM requires a large amount of data. It is difficult to predict fatigue using a small amount of data.
Triantafyllopoulos et al. [327]	Under-represented group (by age range) has lower MAE than well-represented group in the data. Model perform almost the same despite a bias towards female subjects (27:21).
Wang et al. [343]	Min-max normalisation is applied to deal with inter-individual variation.
Zhu et al. [374]	N/A
Chen et al. [71]	Small sample size. Due to the small sample size, 5-fold cross-validation is more appropriate (than 10). Feature optimisation can reduce variance of the model and thus prevent overfitting.
Chen et al. [70]	Sample size was relatively small. Cross-validation was applied to avoid overfitting.
K et al. [181]	N/A
Sadat-Mohammadi et al. [288]	Larger number of subjects to include respiration signals from broader age ranges to increase generalisability of the results.
Wang et al. [344]	Dropout layer can make the model more generalisable to prevent overfitting. Splitting the data into training, validation, and test set ensures generalisability.
Zhang et al. [370]	The lack of a uniform measure of muscle fatigue affects data collection.
Aguirre et al. [4]	Cross-validation addresses the problem of overfitting.
Balaskas and Siozios [28]	Classification varies widely among subjects. Fatigue detected at 43% that proves the model's ability to generalise to unseen running patterns, overcoming the constraint of limited training samples.
Chalitsios et al. [67]	Random Forest is robust against overfitting.
Chen et al. [72]	N/A
Elshafei et al. [108]	Medium-weight dumbbells are the best compromise between recording time and momentum changes as subjects reached fatigue more gradually. The more fatigue data is added to the data, the steeper the decline in ML model performance.
Guan et al. [138]	N/A
Jiang et al. [175]	N/A
Karvekar et al. [189]	Larger number of classes led to a weakening of the ML model performance as more regions with common boundaries to the RPE levels led to more confusion between groups.
Kuschan and Krüger [212]	Results show that SVM classification of fatigue is possible with a small data set.
Lambay et al. [216]	Small size data set does not provide a convincing sample for fatigue prediction. *T3-LOSO* was used as generalised analysis.
Wang and He [341]	N/A
Davidson et al. [95]	Some ML techniques need more input data. CNN require fewer trainable parameters and are more likely to learn useful features from small datasets. Cross-validation to identify a generalisable model. Batch normalisation assists CNN to converge more quickly and improve generalisation.
Luo et al. [238]	Relatively small dataset. Fatigue scores converted to binary labels to reduce intra- and -inter-subject variabilities.
Umer et al. [331]	ML models with larger training set could lead to greater performance. Lower RPEs were harder to predict. Inter-individual variability could be observed under different physical exertion levels. Subjects were only students. Short duration for a single task. Controlled environment.
Wang et al. [345]	SVM based on statistical theory and structural risk minimisation can solve the problems of small samples and overfitting.
Guaitolini et al. [137]	Classifier trained on lower amount of data, having less subjects. Misclassification error could be due to subjects of different ages and training levels. Larger sample size necessary to provide more complete validation.
Maman et al. [242]	Small sample size due to time. Cross-validation is not suitable since the train and test data sets are not independent, for this reason, about 10% of subjects should be left out for cross-validation. 10-fold cross-validation may reduce variation between training and test performance. Bootstrap was applied to reduce bias from model training. Feature reduction was applied to increase generalisability.
Nasirzadeh et al. [261]	Small sample size due to time. Individual variation between participants can skew the results.
Sani et al. [293]	More data sources can lead to more accurate results.

Table B.7 continued from the previous page

Authors	Generalisation / Variance / Small Data / Limitations
Zhang and Wang [364]	Individual differences limit detection results but for the same individual the fatigue characteristics show considerable self-stability.
Chowdhury et al. [77]	The F_1 scores for all folds/users for all classifiers were pooled to increase the statistical power and generalisability of the results.
Geurkink et al. [129]	Low MAE and RMSE results despite small dataset. The relatively small dataset can be used to predict session RPE quite accurately.
Jebelli et al. [171]	Extensive database with a larger number of subjects with more diverse personal characteristics required for future research. Cross-validation was used to confirm that the model generalises.
Karvekar et al. [188]	N/A
Papakostas et al. [265]	Fluctuating model performance for some users. Fatigue detection is challenging due to great variability between self-reports and EMG measurements across subjects and scenarios.
Yang and Ren [357]	N/A
Wu et al. [353]	N/A
Baghdadi et al. [26]	Model performance is assessed with less training data by changing the ratio of training to test sets.
Beéck et al. [33]	29 subjects is a larger data set in sports science than usual. Data collection is time-consuming.
Gordienko et al. [135]	Potential improvement with much bigger data set.
Jamaluddin et al. [169]	N/A
Karthick et al. [187]	sEMG signals have large inter-subject variations.
Aryal et al. [22]	Different environmental factors should be investigated. Task does not reflect actual conditions at work.
Lopez et al. [237]	Transfer learning has the advantage of using much less training data than training the network from scratch. *T3-LOSO* ensures generalisation.
Shahmoradi et al. [302]	The model has difficulty in classifying the fatigue transition state. Subject-independent fatigue recognition should be investigated.
Vandewiele et al. [336]	Rather small data set. Model performance on each of the samples fluctuates a lot. The model can be improved even more by including more variables. Results need to be checked for generalisability for other subjects.
Buckley et al. [55]	N/A
Maman et al. [243]	Models should be investigated, if they are still valid when applied to a larger sample. Fatigue class samples are smaller than non-fatigue samples which is important with small datasets. The data must cover a variety of conditions and should be recorded over an extended period of time for different individuals. Consistent activity conditions for subjects performing the same activity would help avoid variation in the predictive fatigue model.
Carey et al. [65]	A larger dataset would enable a better assessment of the model accuracy. Generalisability for a new player joining the team was not investigated. Cross-validation was not stratified by player identity to ensure out-of-sample predictions, giving a realistic estimation how well the model generalise to new data.
Kupschick et al. [211]	RPE 12 (non-fatigue) and 13 (fatigue) are the classes with the most misclassification. Cross-validation was applied to prevent overfitting.
Pernek et al. [274]	*T3-LOSO* makes model very robust to overfitting problems. Model with different number of subjects was investigated since the model needs a large number of data to scale to real world setting with a variety of different subjects. Prediction error converges after adding the 6th subject.
Bilgin et al. [41]	N/A
Karg et al. [186]	Models are trained individually as the variation in movement changes is large between subjects. No consistent increase in variance and subject fatigue was found. To avoid overfitting caused by a small data set, a filter-based feature selection and removal of highly correlated features is proposed. Few training samples for low and high fatigue for each subject complicate person-dependent ML, so regression is preferred for fatigue detection. Even though model is trained on a small data set, it generalises with unseen data.
Zhang et al. [367]	N/A
Janssen et al. [170]	Relatively homogeneous group. Fundamental movement pattern of walking is not only individual, but also highly situation-dependent, as the characteristics of gait patterns depend on fatigue states.
Subasi and Kiymik [319]	The training set provided to the model was representative of the entire space of interest, so that the trained model had the ability to generalise.
Karg et al. [185]	The modality of expression for several predefined parameters differs among subjects. Significant inter- and intra-individual changes in specific parameters between normal and exhausted human gait.
This thesis	Models do not generalise due to small data. Target group is composed of non-athletes which probably results in a wider variance in the collected data compared to a group of (professional) athletes.

Fatigue Factors Details

C

This section looks in more detail at some of the factors contributing to fatigue that are frequently mentioned in the literature.

Homeostatic factors: Body homeostasis is a self-regulating process by which an organism can maintain internal stability while adapting to changing external conditions [42]. This is achieved by engaging feedback and feedforward pathways that limit variation in one or more control variables [199]. In this context, perceptions of fatigue are likely to contribute to homeostasis by regulating energy expenditure and protecting against overuse injury [199]. Fatigue gradually pushes the body to its limit, breaking the state of homeostasis due to the difference between metabolic energy production or consumption and the accumulation of metabolic waste at the cellular level [108]. Several metabolic stimuli have been proposed to induce muscle fatigue, including muscle glycogen depletion, phosphocreatine, lactate accumulation, low pH, Pi, K1, ammonia, and adenosine triphosphate [199].

Mental factors: Baumeister et al. [30] coined the term ego depletion to refer to a temporary reduction in the self's capacity or willingness to engage in volitional action caused by the prior exercise of volition. However, there is evidence for the metaphor of the human mind as a battery that can be depleted and recharged [114]. However, ego depletion is controversial, Pattyn et al. [269] argue that energy depletion cannot currently be taken as an explanation for fatigue except as a metaphor.

Noakes [263] propose that there must be a central nervous system mechanism, a central governor, that limits further effort to prevent a breakdown in homeostasis. This theory suggests that the brain dynamically and unconsciously modulates the number of active motor units based on a pacing strategy that allows a given task to be

completed in the most efficient manner while maintaining internal homeostasis and a metabolic and physiological reserve [114]. Based on feedback from multiple afferent signals regarding factors such as metabolic rate, fuel reserves, and rate of heat production, a central governor determines subjective feelings of fatigue that increase as the estimated limits of homeostatic stability are approached, ultimately leading to a reduction in motor recruitment and cessation of exercise [114]. Overriding the central governor can lead to physical injury, such as torn muscles, ruptured tendons, and broken teeth [114].

Psychological factors: Psychological factors that contribute to the perception of fatigue include perceptions of effort, expectations, familiarity, motivation, temporal and performance feedback, arousal, and mood [199]. Individuals must continuously self-regulate different affective states induced by different perceptions (e.g., effort perception, exercise-induced pain perception), thoughts (e.g., related to task-termination or distractors), and behaviours (e.g., stopping the task or increasing the effort), with consequences for their motor performance [34]. For example, muscle fatigue from brief anaerobic exertion, such as sprinting, can be moderated by several psychological factors, including hypnosis, sudden noise, music, and deception about workload, suggesting control by central mechanisms rather than, or in addition to, peripheral energy stores [114].

Studies have shown that orally swishing glucose without ingestion leads to improvements in exercise performance and activation of brain areas associated with reward and motivation. The extent of fatigue is determined by the expected levels of glucose relative to the estimate of its anticipated need, combined with the expected levels and needs of other resources that may be required for successful goal completion. Muscle fatigue begins almost immediately after the onset of exercise; it is highly unlikely that any substrate could be depleted so quickly. These studies suggest that fatigue and performance limits are mediated by the central nervous system [114].

The ability of psychological factors such as beliefs and motivation to overcome self-regulatory fatigue may be due to a the expectation of a disturbance in homeostasis, the lack of expected benefit to the desired goal, or both. In laboratory studies of self-regulatory fatigue, the motivation to perform a task is usually extrinsic: an experimenter asks participants to do their best at a task that is irrelevant to the participant. In these cases, even seemingly simple tasks can cause considerable fatigue due to low levels of motivation and perceived benefit. Tasks that are intrinsically motivated may produce less self-regulatory fatigue [114].

Central factors: Fatigue during both motor and cognitive tasks is additionally driven by central nervous system mechanisms. Neurophysiological studies in healthy human subjects show changes in motor cortex and spinal excitability associated with fatigue during motor tasks and suggest that deficits in central drive account for a significant percentage of fatigue depending on task demands [199].

Signals such as the level of muscle recruitment are transmitted through the spinal cord to the thalamus, a kind of relay station for sensory and motor signals from the periphery to the cerebral cortex. Other signals to the brain probably travel a different route. The vagus nerve, which is 80-90% afferent fibres, carries a vast amount of information from the body to the brain, including the state of the heart, lungs, gut, and immune system [114].

Peripheral factors: Peripheral nerves and physiological changes in muscle contribute to fatigue in healthy humans and can be referred to as peripheral factors, based on anatomical distinctions between the peripheral nervous system and the central nervous system. Alternatively, the terms “contractile factors” and “activating factors” can be used to distinguish between peripheral mechanisms within the muscle and those that provide the activating signal [199]. However, the mechanisms of physical (motor) performance fatigue are not fully understood: data suggest that changes in the nervous system and muscle during motor tasks contribute to the decline in motor performance [34]. Other contributors may include the type, amount

and intensity of physical work and effort, as well as neuromuscular characteristics, metabolite storage and buffering capacity [269].

The degree of motor fatigue may differ between women and men during fatiguing isometric and dynamic tasks. Males usually show greater motor fatigue than females during single-joint isometric and slow to moderate velocity muscle actions and whole-body exercise. The sex difference in motor fatigue is reduced in fast-paced muscle actions and is highly dependent on the muscle(s) tested. Motor function declines with age due to structural and functional changes within the neuromuscular system: Age-related differences are specific to the muscle group studied [34].

One of the most important and studied factors influencing the extent of motor performance fatigue is the characteristics of the motor task, which determine the stress imposed on the involved physiological subsystems. The magnitude of the decline in maximal voluntary force and the relative contribution of changes in muscle activation and contractile function are strongly dependent on the duration and intensity of the exercise, the mode and velocity of muscle action, and the engaged muscle mass [34].

Pathological factors: Pathological changes in the peripheral nervous system and muscle may also influence fatigue [199]. For example, illnesses and diseases such as myasthenia gravis, a chronic autoimmune disease, impair neuromuscular connectivity and result in decreased muscle strength and endurance during repetitive muscle activity [108]. However, the degree of actual physical fatigue is sometimes difficult to measure, especially in cases where a pathology is present [50].

Age factors: Fatigue is a common condition affecting people of all ages, but it appears to be particularly prevalent in older people. While demographic factors play a role, the exact causes and effects of fatigue are still being investigated [217].

Sleep factors: Sleep deprivation can severely affect performance, motivation, perception of exertion, cognition and many physiological functions [143]. Elsais et al. [107] report strong correlations between fatigue and thermoregulation, as well as

sleep quality and orthostatism (a condition caused by low blood pressure). A related effect is circadian rhythms, which can also influence fatigue levels [248].

Time factors: Perceived fatigue can be measured at rest or during physical activity, whereas performance fatigue is quantified as the rate of change in a criterion outcome due the adjustments made during a fatiguing task [110].

Domain factors: Accurate assessment of fatigability or momentary perception of fatigue requires specification of the performance domain and task. Performance domains include sustained contractions, repetitive movements, skilled sequences, cognitive functions (such as working memory and attention) and verbal abilities. While some factors, such as arousal, affect performance across domains, many factors are domain-specific. For example, motor and cognitive tasks induce fatigue at different rates and stress different physiological factors, but even within well-defined motor tasks, subtle differences in performance strategy may influence the rate of fatigability [199].

Fatigue Exercise Load

D

A concept related to fatigue is exercise load, which can be interpreted in three ways: Firstly, by physical measures such as power, work and energy, torque or velocity. Secondly, physiologically in absolute terms such as VO_2max or by relative values such as heart rate. Thirdly, in terms of ratings of subjective intensity as perceived by the subject [50].

Fatigue is an inevitable consequence of incremental tasks such as training exercises [108]. Exercise-induced fatigue occurs when the effort required by the exercise task equals the maximum effort the subject is willing to exert to succeed in the task, or when the subject believes they have exerted a true maximum effort and continuation of the exercise is perceived as impossible [269]. To explain the variance between subjects and workloads, different physiological variables, and their interactions need to be measured and weighted to develop a model for prediction [50].

Muscle power output decreases as exercise duration increases, following a hyperbolic relationship characterised by critical power. This is the maximum sustainable power output without fatigue accumulation. Above critical power, a limited amount of additional work can be performed before exhaustion, typically within 30 minutes. Critical power marks the transition from heavy to severe exercise intensity where physiological adaptations can no longer maintain homeostasis [110].

Halson [143] distinguishes between external and internal load. External load is defined as the task performed by the subject, measured independently of their internal characteristics (e.g., power output, speed, acceleration, time movement, neuromuscular function). Internal load, or the relative physiological and psychological stress imposed, is also critical in determining training load and subsequent adaptation (e.g., heart rate to RPE ratio, heart rate variability, heart rate recov-

ery, TRIMP, lactate concentration, lactate to RPE ratio, biochemical assessment, questionnaires, psychomotor speed, and sleep).

In addition, the validity of a measure of a load indicator depends on the context. For example, heart rate is a valid measure of internal load for endurance training but not for resistance training. Even within the same context, a single measure of exertion may not have the same level of validity. Heart rate is a less valid indicator of internal load for short duration, intermittent high intensity efforts compared to long distance or interval training. Measures of external load are specific to the type of training performed. Muscle fatigue increases both heart rate and perceived exertion, whereas mental fatigue increases only perceived exertion [164].

The problem is finding the right level of exercise intensity: when designing an experiment, researchers need to find a compromise between recording time and fatigue-inducing task momentum, so that the majority of subjects reach fatigue more gradually. Moreover, it is important to ensure that factors causally related to fatigue can be distinguished from compensatory or other associated factors, which may also change over the course of task performance [199]. Causal factors should show correlations with performance decline rather than change over time [199].

E

Fatigue in the Context of Stress

Attention is drawn to the field of stress in the context of sports, its relation to fatigue, and how studies have attempted to measure it during exercises. There is a long debate across multidisciplinary fields about the concept of stress [111]. Since each discipline has its own concepts on stress, a common definition is unlikely [23, 125]. A common understanding exists, but the meaning depends on one's concept [125]:

- Stimulus (1939) by Cannon (physiologist): Stress is associated with maintaining homeostasis and induced by the environment "stressors", e.g., heat, diseases, exhaustion, demands, or submission dead-lines [64].
- Response (1956) by Selye (endocrinologist): Stress is how an individual or organism reacts to events or conditions, e.g., anxiety or fatigue [298].
- Transactional (1984) by Lazarus (psychologist): Stress is an imbalance between demands placed on a person, and opportunities as well as resources available to meet those demands ("coping") [223].

Stress can be classified as acute or chronic [111, 151]. While chronic stress is pathological and psychological in nature, acute stress is the immediate response of the body to a stimulus (stressor) [151, 298]. The acute response triggers alertness, energy release, physiological regulation, and immunological activation to compensate for the effects of the stressor [151]. During training exercises, the body experiences an acute stress response in which more oxygen and energy are required. The heart rate increases so that more blood is pumped through the body and thus oxygen is transported to improve cardiorespiratory function [29]. Stress could be understood as a response to a disturbance of homeostatic balance by events

or conditions (stressors) [298]. For example, untrained people suffer from more stress due to higher demand for oxygen and energy, while trained people become accustomed to use less oxygen; their body will eventually feel the stress over a longer period [29, 87]. The physiological reactions are summarised as follows:

- Sympathoadrenal system (SAM axis): Sympathetic activation and parasympathetic withdrawal cause increased heart rate and respiratory rate, bronchial and pupil dilation, sweaty skin, and other symptoms. The body is rapidly prepared for a physical "fight or flight" stress response [29].
- Hypothalamic-pituitary-adrenal axis (HPA axis): Slowly activated by the secretion of cortisol leading to increased catabolism, anabolism inhibition, and depression of the immune system. Typically activated by mental tasks [133].

As stress has different physiological sources and effects, a single stress marker cannot holistically assess the stress response of a person [11]. A method that considers multiple stress response reactions is needed. For example, skin conductance can be an indicator of sympathetic nervous system activation (SAM axis), while electrocardiogram can detect the activation of the HPA axis and SAM axis [3].

In addition, stress is highly subjective and individual in all aspects [130, 132]. There is a lack of research on methodological and measurement standards to determine stress during challenging contexts such as training exercises [23, 3, 130, 213, 299], for which stress is a natural physiological response [29, 317, 18, 141]. In principle, there are countless stimuli that are associated with stress [132, 264]. One of them is the performed quality of a training exercise [8, 156]. Fatigue is another stimulus for stress [200, 119]. Physical activity could be viewed as providing stimuli that promote specific and varied adaptations of the body depending on the type, intensity, and duration of exercise performed [87, 141, 142]. Chronic exercise training does not eliminate the acute exercise response, but it can attenuate the overall effect of the response as the body adapts to the training stimulus in a positive way. An excessive intensity and/or volume of training may lead to maladaptation

[141, 142]. Hence, a stress response is dependent on the athlete and the exercise. An unfamiliar exercise is likely to elicit a higher metabolic stress response than a familiar, routine exercise, e.g., a long-distance runner will probably have a different stress response profile for a given exercise than a weightlifter. Exercises represent an effective methodological tool to study the body's response to metabolic stress, and from a clinical perspective, offers an alternative treatment choice to drug intervention strategies [29].

Thus far, only a few studies have attempted to investigate physiological stress during training exercises. For example, [241] focused on the effect of physical and mental stress on the heart rate as well as cortisol and lactate concentrations. They found that the heart rate is most sensitive to physical and mental stress. [156] investigated the influence of physical activity on stress recognition with physiological responses. The authors used different stressors to induce stress and found that, among others, stress models for each physical activity should be built due to variations in physiological changes caused by physical activity. Alamudun et al. [8] introduced two multivariate signal processing algorithms to cope with the differences in physiology between participants and changes in physical activity. They found that these two algorithms can bring noticeable improvements for the process of stress prediction. Wong et al. [351] used IMU data to distinguish stress and high intensity activity in daily life.

Tab. E.1.: An overview of commonly used stress markers.

Subjective Stress Markers	**Objective Stress Markers**	
E.g., interviews, self-reports, and questionnaires.	Obtrusive	Unobtrusive (real-time)
	Salvia, hair, and blood samples (e.g., cortisol or lactate).	Wearables (e.g., heart rate), contextual (e.g., air quality), video-based (e.g., thermal imaging), behavioral (e.g., physical activity).

Based on our literature work, we created a tabular overview of existing stress markers (see Table E.1). Stress (and fatigue) markers can be classified as subjective or objective depending on the measurement technique [136]. Subjective stress markers, on the one hand, are traditionally used by psychologists in the form of ques-

tionnaires, interviews, or self-reports, which are usually conducted retrospectively. Subjective markers are not suitable to continuously monitor stress during training exercises but can be used to determine stress levels before and after an exercise. Objective stress markers, on the other hand, are quantifiable and cover physiological, physical, behavioural responses, and other contextual data. They can reduce the possibility of self-deception, falsification, fabrication, attention, or recall bias, which is usually present in subjective markers [307]. Objective markers are measured either obtrusively or unobtrusively [23]. Biomedical researchers rely on obtrusive biochemical markers, typically hormones, to measure stress [23, 317]. One of these hormones is cortisol, which is commonly used in studies on stress [130]. Another less expensive marker is lactate [80, 153] which was once incorrectly attributed to muscle fatigue [80]. Such obtrusive biochemical markers provide accurate quantitative data [23]. However, they are not suitable for real-time monitoring systems due to their inherent nature and that they, at times, necessitate analysing data in a laboratory. Unobtrusive stress markers, such as heart rate or muscle activity, are measured by sensors that are attached to the body. They provide continuous data in real-time and do not require analysis in a laboratory [130]. Yet, unobtrusive markers are susceptible to noise or artefacts due to individual's body parts movements or activities [130]; however, studies show that they can provide relevant indicators to determine stress [23, 3, 130, 132, 136, 159, 299, 351, 362].

Regardless of the stress marker, Arza et al. [23] state that a single stress marker cannot globally assess an individual's stress response, because stress causes different physiological reactions, and a multi-variable approach is therefore suggested. Due to the multifaceted characteristics of stress, determining a ground truth is a difficult process [130]. Some studies use subjective measures of perceived stress. Other studies rely on biosignals or biomarkers that they consider reliable for determining stress. In many studies, ground truth is established by placing a subject in a neutral and in a stressful situation to label the collected data accordingly. Others use the

amount of workload and cognitive demand that is being applied as the stressor [23, 130, 144].

In summary, stress cannot be objectively and unobtrusively monitored in real-time [132]. Determining stress is challenging because of the subjectivity and individual nature of stress [132]. Moreover, the start, the duration, and the intensity of a stress event is often not clearly identifiable [132]. There is also no commonly agreed methodological or measurement standard for unobtrusive markers [23, 130]. The relationship between the body's activation of biochemical stress markers and the intensity of the stress perceived is both complex and understudied [23]. However, it has been shown that unobtrusive stress markers can be used to approximate stress (and implicitly fatigue) in real-time [23, 130].

RPE Scale Instructions

F

Scale Instructions

While exercising we want you to rate your perception of exertion, i.e., how heavy and strenuous the exercise feels to you. The perception of exertion depends mainly on the strain and fatigue in your muscles and on your feeling of breathlessness or aches in the chest.

Look at this rating scale; we want you to use this scale from 6 to 20, where 6 means "no exertion at all" and 20 means "maximal exertion."

6	No exertion at all
7	
8	Extremely light
9	Very light
10	
11	Light
12	
13	Somewhat hard
14	
15	Hard (heavy)
16	
17	Very hard
18	
19	Extremely hard
20	Maximal exertion

Borg RPE scale
© Gunnar Borg, 1970, 1985, 1994, 1998

Figure 7.1 The Borg RPE scale for perceived exertion.

9 corresponds to "very light" exercise. For a normal, healthy person it is like walking slowly at his or her own pace for some minutes.

13 on the scale is "somewhat hard" exercise, but it still feels OK to continue.

17 "very hard" is very strenuous. A healthy person can still go on, but he or she really has to push him- or herself. It feels very heavy, and the person is very tired.

19 on the scale is an extremely strenuous exercise level. For most people this is the most strenuous exercise they have ever experienced.

Try to appraise your feeling of exertion as honestly as possible, without thinking about what the actual physical load is. Don't underestimate it, but don't overestimate it either. It's your own feeling of effort and exertion that's important, not how it compares to other people's. What other people think is not important either. Look at the scale and the expressions and then give a number.

Any questions?

Fig. F.1.: BORG Scale instructions by Borg [50].

RPE Principles

G

Studies using RPE should consider the following principles in its application [50].

The first principle is to explain to the subject the purpose of the study. While the subject may naturally expect the investigator to ask questions about perceived exertion during a test, a brief rationale for the need for such questions should be given. The explanation need not be elaborate; a simple statement can convey the meaning.

The second principle is to provide clear instructions as to what is to be assessed (see also Figure F.1 in the Appendix). It is essential to provide comprehensive instructions and explanations on how to assess perceived effort. The subject should understand that the focus is on their subjective perception. It is essential that the subject focuses their attention inward, concentrating on their internal subjective feelings, rather than dwelling on the physical tasks or physiological cues and responses. The aim is for the subject to approach the assessment with spontaneity and a 'naive' attitude, taking an introspective rather than a stimulus-orientated approach. The subject should refrain from considering external opinions or hypothetical responses in similar situations and rely on their own feelings. To ensure accurate and reliable results, a clear and concise scale instruction for the rating process is necessary. This instruction should strike a balance between simplicity and comprehensiveness, avoiding unnecessary brevity or complexity. It should include an explanation of the aspects to be rated, how the scale works and the importance of verbal anchors. In addition, the instructions should outline the expectations of the respondent and emphasise the importance of being a conscientious rater. The exclusivity of the anchors specified in the instructions and the verbal descriptors on

the scale is crucial. After the instructions have been given, it is important that the experimenter encourages the subject to ask questions.

The third principle considers the contextual elements of 'where and when' Control of the physical environment can be challenging, with social factors introducing potential confounds. Variables such as unfamiliar equipment, intimidating instruments, ambient music or noise, temperature fluctuations and the presence of other people engaged in other activities can all have a significant impact on both performance and subjective scoring. In cases where control is not possible, it is imperative that these elements are recognised and taken into account in the interpretation of results. Although many laboratory settings allow for strict control of confounding factors, it is important to maintain optimal conditions for RPE testing when assessing work capacity. Individuals undergoing testing should be well rested, alert and following their typical daily routine. Testing immediately after eating or taking medication (unless necessary) is discouraged. In addition, psychological factors play a role in the perception of exertion, emphasising the importance of choosing a time for testing when the individual is calm, relaxed and ideally feels a sense of control over their situation.

The fourth principle is about how to assess and evaluate. A positive relationship and co-operation should be established between the investigator and the subject. The study should not only be objectively controlled, but also standardised to ensure consistent understanding and responses across subjects. Whilst objectivity is maintained through standardisation, the experimenter must have the skills to be flexible and accommodate potential changes due to individual personality factors and unexpected situational elements. If necessary, interruptions during the test, accompanied by clarification of instructions, may be required. Appropriate management of emotional factors and health problems should be addressed as necessary. Athletes, in particular, may tend to underestimate their perceived exertion, driven by a desire to demonstrate high levels of fitness. Conversely, people with low motivation may either exaggerate or downplay their exertion levels based on personal

motivations. In cases where collaboration is hindered by misunderstanding, fear or low motivation, it is essential to conduct a detailed interview. The purpose of this interview is to emphasise the importance of following instructions, to promote a more cooperative testing environment and to ensure the reliability of the results.

The fifth principle is to establish a robust protocol for systematically documenting responses during the test. Particular attention should be paid to comprehensive observations, covering both behavioural and physiological aspects, as these may prove crucial in subsequent evaluations. To reduce the risk of data entry errors, a verification protocol should be followed for recording responses and other reactions during a test.

RPE Interpolation

H

In the case study, RPE were collected as labels at 10 second intervals, resulting in multiple squat repetitions within each interval sharing the same RPE. As RPE is known to change linearly with exercise intensity [33, 113, 51], the RPE for each repetition can be interpolated by calculating an overall slope via linear regression [4]. To do this, all RPE collected for each set must be normalised so that they start from a common baseline (e.g., zero). A linear regression model, such as the MATLAB function fitlm[1], can then be used to analyse the baseline RPE and compute an overall slope value.

Figure H.1 illustrates how the slope of 9.6528 was determined across all subjects. The slope obtained from the linear regression was then used to interpolate RPE labels for each individual squat repetition based on either the first or last RPE in a set. It was also necessary to constrain the interpolated values within the upper and lower limits of the RPE scale to ensure their validity. Table H.1 shows an example of how interpolated RPE values are stored in a segment file.

[1] `https://de.mathworks.com/help/stats/fitlm.html`

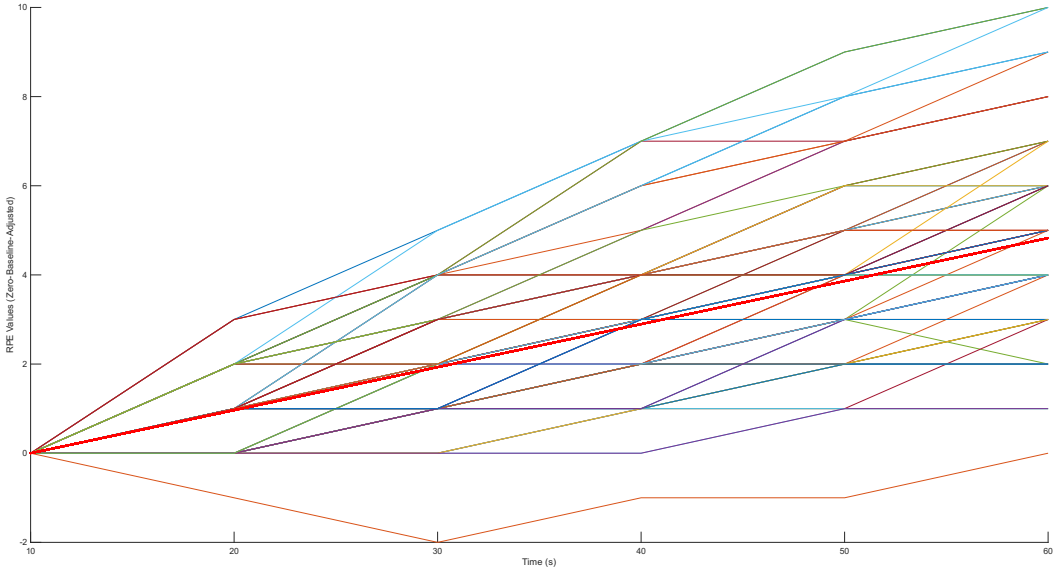


Fig. H.1.: The RPE for each set were normalised to establish a common zero-baseline, facilitating the computation of linear regression to determine the overall slope (thick red line).

Tab. H.1.: Example of segments including raw and interpolated RPE.

SegmentNumber	Exercise	Set	SegmentStart_ms	SegmentEnd_ms	Label	Interpolated
1	Squats	1	7753.3	9753.3	11	10
2	Squats	1	9753.3	11853.3	11	10
3	Squats	1	11853.3	13920.0	11	11
4	Squats	1	13920.0	15920.0	11	11
5	Squats	1	15920.0	17953.3	11	11
6	Squats	1	17953.3	20086.6	12	11
7	Squats	1	20086.6	22120.0	12	11
8	Squats	1	22120.0	24153.3	12	12
9	Squats	1	24153.3	26086.6	12	12
10	Squats	1	26086.6	28086.6	12	12
11	Squats	1	28086.6	30220.0	13	12
12	Squats	1	30220.0	32320.0	13	12
13	Squats	1	32320.0	34353.3	13	13
14	Squats	1	34353.3	36386.6	13	13
15	Squats	1	36386.6	38486.6	13	13
16	Squats	1	38486.6	40553.3	14	13
17	Squats	1	40553.3	42553.3	14	13
18	Squats	1	42553.3	44553.3	14	13
19	Squats	1	44553.3	46553.3	14	13
20	Squats	1	46553.3	48586.6	14	14
21	Squats	1	48586.6	50653.3	14	14
22	Squats	1	50653.3	52720.0	14	14
23	Squats	1	52720.0	54753.3	14	14
24	Squats	1	54753.3	56753.3	14	14
25	Squats	1	56753.3	58786.6	14	14
26	Squats	1	58786.6	60853.3	13	14
27	Squats	1	60853.3	62786.6	13	14
28	Squats	1	62786.6	64753.3	13	13
29	Squats	1	64753.3	66753.3	13	13
30	Squats	1	66753.3	68820.0	13	13
1	Squats	2	137386.6	139253.3	11	10

RPE vs Heart Rate

There is a strong correlation between heart rate and RPE [45, 50]. However, preliminary tests with the heart rate sensor have shown that heart rate is highly dependent on the order of the exercises and therefore does not always correlate with RPE. Figure I.1 shows the recorded heart rate as a blue line during six different training exercises. The blue numbers represent the RPE.

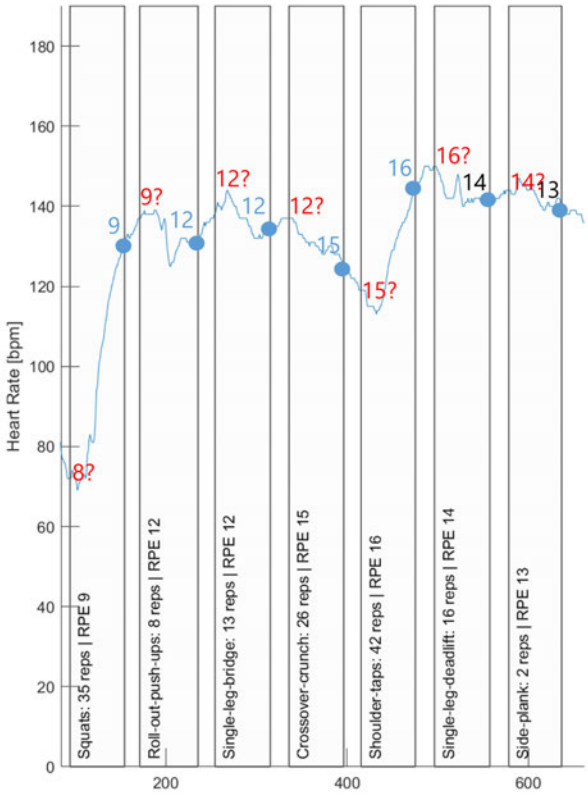


Fig. I.1.: RPE compared to heart rate progression during six consecutive training exercises.

Figure I.2 shows a sketch of the general trend of heart rate progression during the six consecutive training exercises for all subjects.

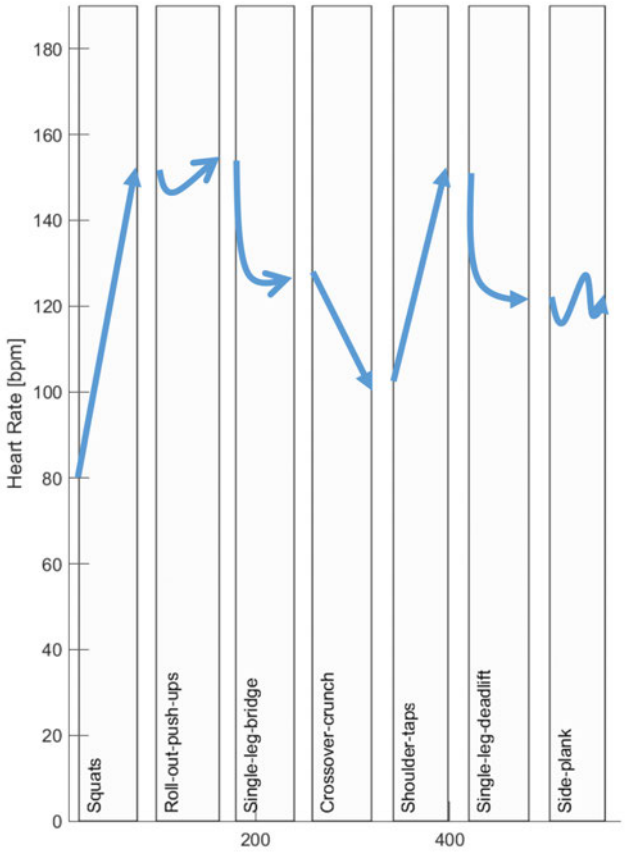


Fig. I.2.: Sketch of the general trend of heart rate progression during six consecutive training exercises for 20 subjects.

HAR Sensing Techniques

J

HAR sensing techniques by Bian et al. [39] based on [349, 124].

Field Sensing The field is a concept in physics that refers to an area in which any point is affected by gravitational, magnetic, or electrical force. Gravitational field-based HAR tasks mainly use the pressure sensed by pressure sensors caused by the weight of the body. Magnetic field strength can be sensed by magnetometers and is often included in IMUs. Electric field-based HAR applications can be active or passive. An active electric field-based HAR application delivers the field variation as a signal source when the field is emitted by the environment and the human acts as an intruder. A passive one provides the field variation when considering the electric field emitted from the body itself to the ground, as the human body is a conductor and can store the charges.

Physiological Sensing Physiological sensing refers to natural physiological and kinematic signals activated by an organism. Physiological sensing can include blood sampling, blood pressure, heart rate, respiration, phonation, muscle and joint movement, or facial expression. A subfield is electrophysiology, which focuses on the electrical properties of neurons, molecular and cellular substances. These can be monitored using various (wearable) sensors, such as electromyography, electrocardiogram, electroencephalogram, and electrooculography. A more detailed description of human bioelectric signals can be found in Shen et al. [305].

Mechanical Kinematic Sensing Mechanical sensing refers to the mechanical mobility and deformation when a force is applied to or from the target. The mobility and deformation are sensed by mechanical sensors, which convert the mechanical change into electrical signals. Mechanical sensors are widely used to monitor body activity,

such as kinematic sensors, which measure motion characteristics such as velocity, acceleration and rotation. Kinematic sensors, such as inertial measurement units (IMUs), have become a prominent sensing approach in industrial applications and scientific research. IMUs typically contain multiple sensors such as accelerometers, gyroscopes, and magnetometers.

Wave Sensing Wave sensing is a non-contact sensing technique based on the propagation properties of waves. Three types of wave sensing approaches are mainly used for HAR tasks: The first is radio frequency, with frequencies ranging from 3 kHz to 300 GHz, such as WiFi, Bluetooth, mmWave or ultra-wideband. The second is acoustic, a mechanical wave including vibration, sound, ultrasound and infrasound. The third is optical, an electromagnetic signal with the typical extremely high frequency in the THz range (e.g. image or video).

Hybrid / Others Other techniques are human body capacitance and infrared. Human body capacitance is a biological variable that describes the capacitance between the human body and the environment. Infrared is electromagnetic radiation with wavelengths longer than visible light and manifests as heat energy from objects with temperatures above absolute zero, commonly measured by passive infrared sensors or thermographic cameras. Finally, there are hybrid techniques, which can be any combination of the above.

K

Markerless vs Marker-based Motion Tracking

Marker-based infrared tracking (AR tracking) was used to verify the accuracy of the PE data. Figure K.1 shows both markerless PE and marker-based AR tracking signals during squats for the y-axis. The AR tracking signals were normalised to match the range of the PE. Both signals are quite similar and synchronised.

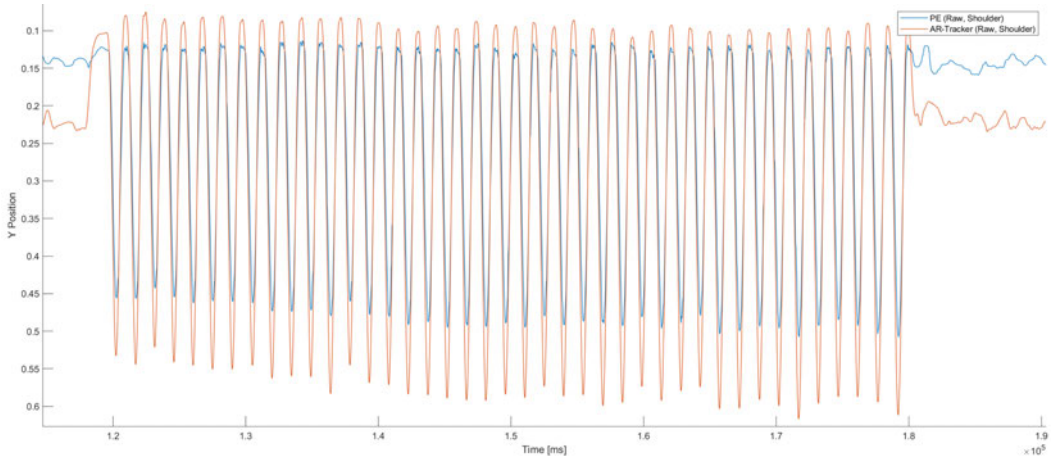


Fig. K.1.: Comparison of PE and normalised AR tracking signals for the y-axis during squats.

Filter Methods and Phase Shift

A Butterworth filter is a IIR filter designed to have a maximum flat frequency response in the passband, i.e., it does not introduce waves or distortion. It is commonly used in applications where smooth filtering is required, such as audio processing or control systems. The key characteristic of a Butterworth filter is that it provides a smooth transition between the passband (where the signal passes through) and the stopband (where the signal is attenuated). Unlike some filters, which can oscillate or have waves, a Butterworth filter is designed to be flat and non-oscillating in the passband, which reduces distortion of the filtered signal [76, 106, 314, 266]. A Butterworth filter uses poles in the complex plane, arranged to give the smoothest possible transition. The filter's transfer function defines how the frequency components of the input signal are modified, with lower frequencies passing through and higher frequencies being attenuated, depending on the filter's design. The order of the Butterworth filter affects the sharpness of the roll-off. A higher order filter will have a steeper slope after the cut-off frequency, but will require more complex design and implementation [106, 266, 314].

The general Butterworth filter formula for a low-pass filter is:

$$|H(\omega)| = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}$$

Where:

- $|H(\omega)|$ represents the magnitude of the frequency response at a given frequency (normalised gain),

- ω is the frequency (of the input signal),
- ω_c is the cut-off frequency,
- n is the filter order.

The higher the order n, the steeper the transition between the passband and stop-band.

The discrete-time Butterworth filter transfer function is:

$$H(z) = \frac{b_1 + b_2 z^{-1} + b_3 z^{-2} + \cdots + b_r z^{(r-1)}}{a_1 + a_2 z^{-1} + a_3 z^{-2} + \cdots + a_r z^{(r-1)}}$$

Where:

- $b_1, b_2, ..., b_r$ are the numerator (feedforward) coefficients,
- $a_1, a_2, ..., a_r$ are the denominator (feedback) coefficients,
- z represents a delay operator in discrete-time.

Given the filter order n, the function returns b and a with r samples, where $r = n + 1$ for low-pass and high-pass filters and $r = 2 * n + 1$ for bandpass and bandstop filters.

Figure L.1 compares IMU signals filtered by different filtering methods. Some filters caused a phase shift, i.e., the signal was shifted to the right (delayed) [76].

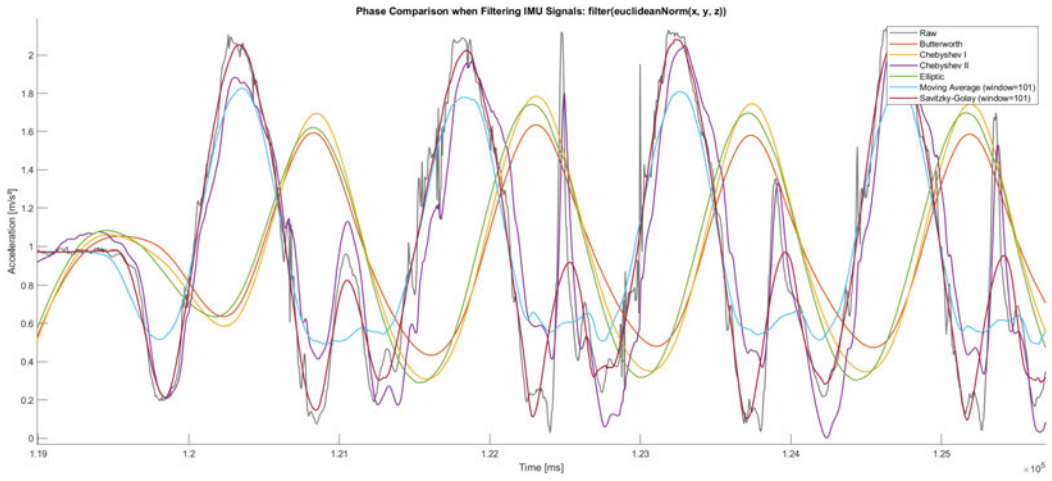


Fig. L.1.: Comparison of different filtering methods in regard to phase shift.

Research Onion

M

The research onion by Saunders [295] which is part of defining the research design.

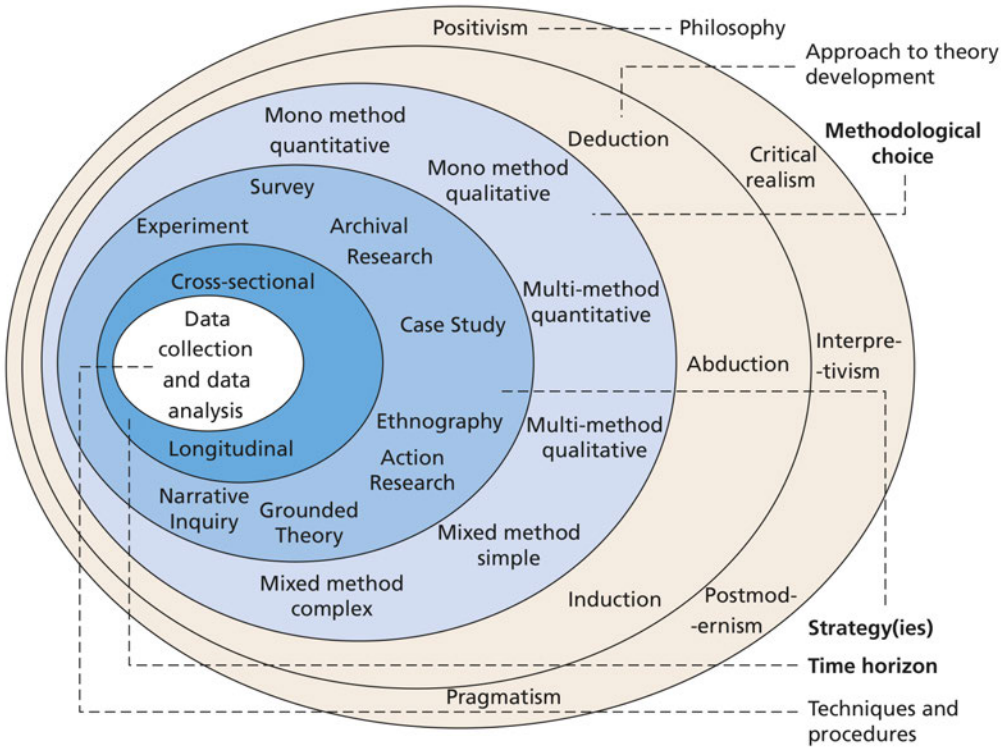


Fig. M.1.: The research onion by Saunders [295].

MediaPipe Pose

N

MediaPipe Pose leverages a CNN architecture similar to MobileNetV2 [292], tailored for on-device, real-time fitness applications. A CNN for pose estimation typically consists of two parts: an off-the-shelf generic pre-trained network such as ResNet, AlexNet, Cascaded Pyramid, Hourglass, HRNet, or MobileNetV2 to extract features (also known as the backbone network), and the prediction head that predicts human poses with the extracted features.

Rather than using backbones designed for classification tasks directly, it's important to refine them specifically for human PE. Regarding prediction heads, there are mainly two representative solutions: directly predicting joint coordinates, which is considered as the regression paradigm, or generating an intermediate heatmap representation before computing joint coordinates. In the regression paradigm, fully connected layers are typically employed to determine precise key point coordinates. In the heatmap prediction paradigm, the process of upsampling is commonly used to produce higher resolution heatmaps [220].

The CNN in MediaPipe Pose is a variation of the open-source BlazePose model, which utilises heatmaps and regression to key point coordinates. The heatmaps are generated by running an image through multiple resolution banks in parallel to capture features at different scales simultaneously. It's worth noting that heatmaps and offset loss are only used during the training phase and are then removed from the model before inference. The heatmaps monitor the lightweight embedding, which is then used by the regression encoder network [31] (see Figure N.1).

This approach is partly inspired by the design of stacked hourglass networks, where multiple hourglass networks are stacked end-to-end, with the output of one serving as the input to the next. This provides the network with a mechanism

for repeated bottom-up (low resolution) and top-down inference, allowing initial estimates and features to be re-evaluated across the entire image. Each hourglass module integrates both local and global cues. Asking the network to produce early predictions requires a high level of understanding of the image even before the full network has been traversed. Subsequent stages of bottom-up and top-down processing allow for a more thorough reconsideration of these features [262].

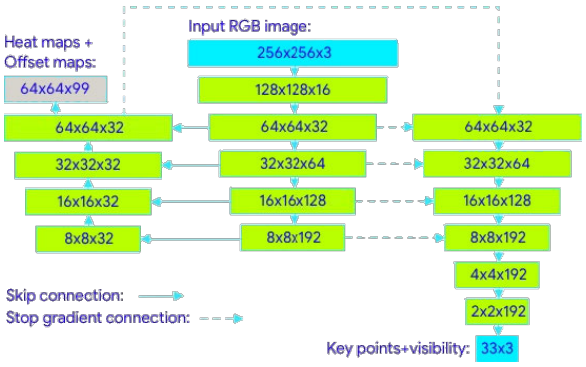


Fig. N.1.: Network architecture of BlazePose [31].

This back and forth between scales is particularly important because preserving the spatial location of features is essential for the final localisation step. The precise position serves as a critical cue for other decisions made by the network. In a structured problem like pose estimation, the output results from the interplay of many features that should converge to form a coherent understanding of the scene. Any conflicting evidence or anatomical impossibilities act as significant indicators that an error may have occurred somewhere along the way. By iterating through these scales, the network can retain accurate local information while continually assessing and reassessing the overall coherence of the features [262].

In addition, skip-connections are incorporated across all stages of the network to achieve a balance between high and low-level features. The gradients from the regression encoder are not propagated back to the heatmap trained features. The model is trained to predict body pose in relative coordinates of a metric space with origin at the centre of the subject's hips [31].

IMU Signals

O

In the following, IMU signals from different subjects are highlighted where a noticeable change in the signal took place during the squat exercise. All IMU signals were filtered using a Butterworth filter. Signal signature refers to the number of local minima and maxima for a repetition.

Figure O.1 shows the IMU signals of a subject where the signal signature of the performed squat training changes within the same set.

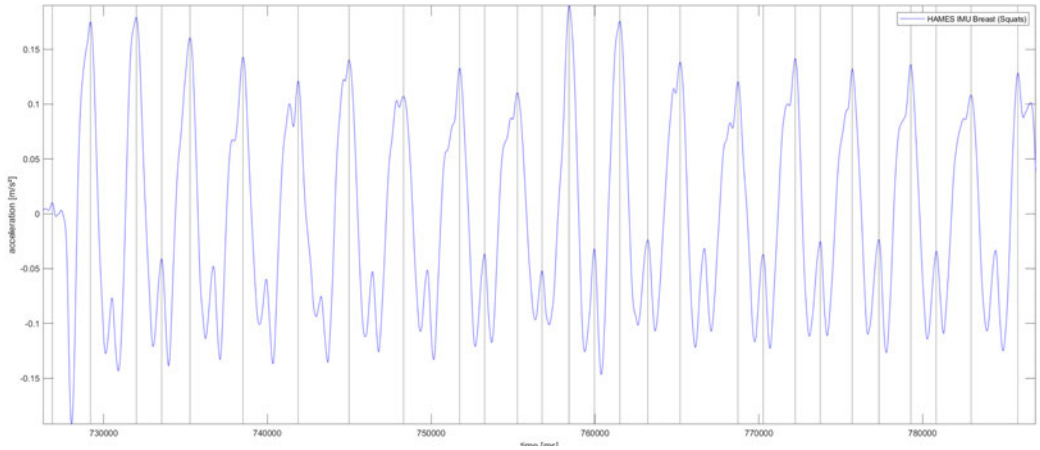


Fig. O.1.: Same subject, but the IMU signal signature changes within the same set.

Figure O.2 shows the IMU signals of a subject where the signal signature of the performed squat training changes between the first and third sets.

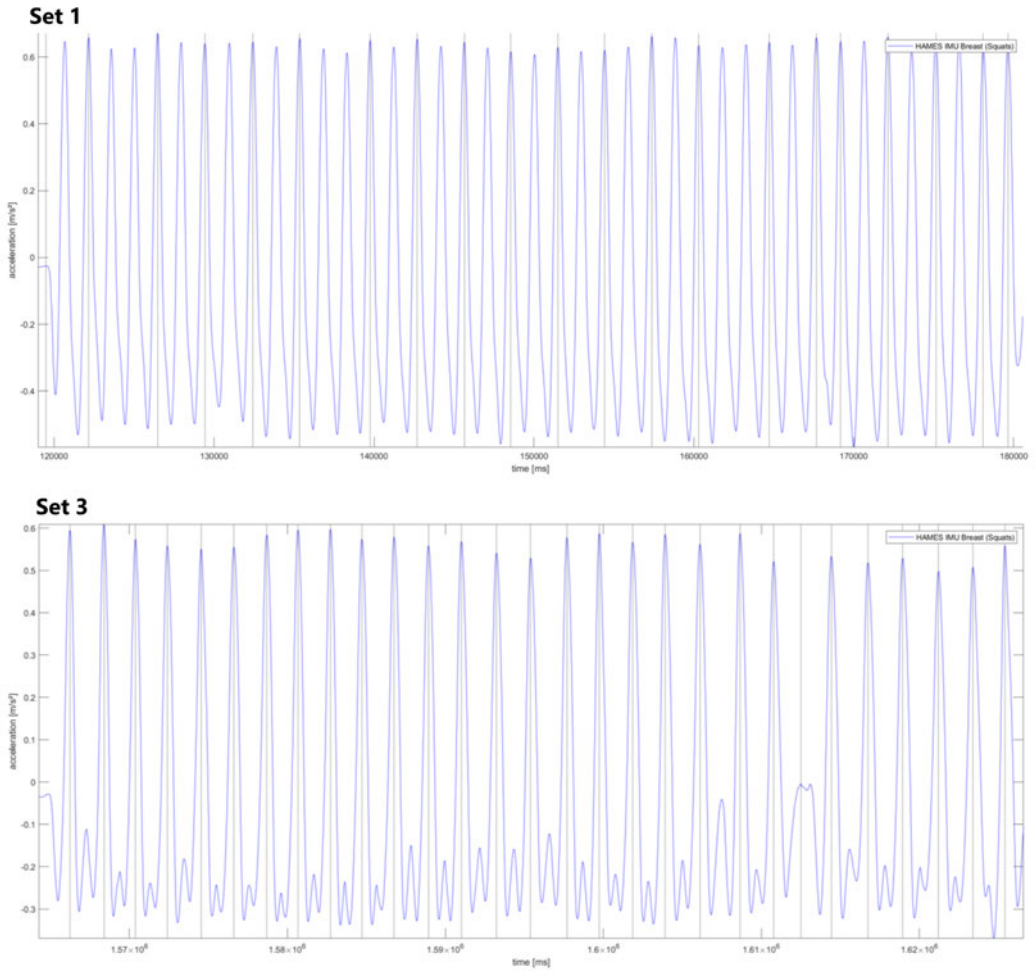


Fig. O.2.: Same subject, but the IMU signal signature changes between the first and third sets.

Figure O.3 shows the IMU signals of a subject where the signal amplitude of the performed squat training changes within the same set.

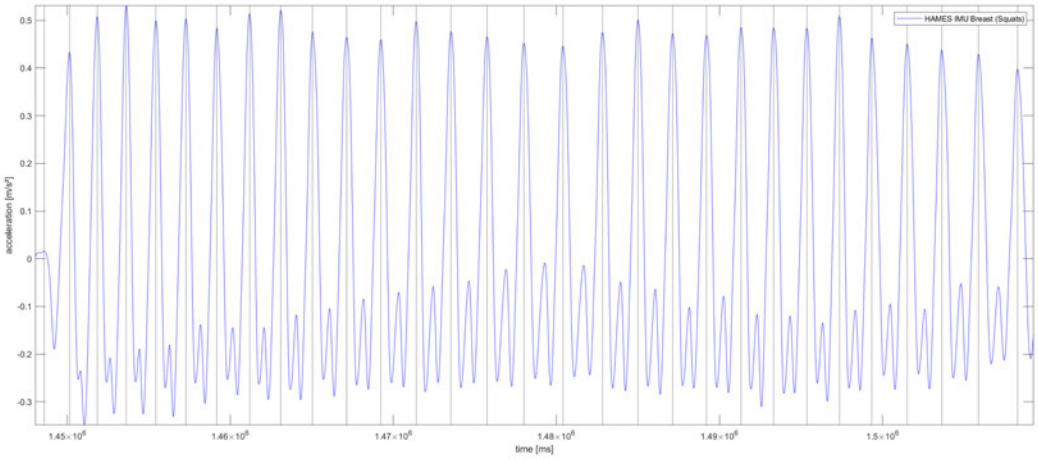


Fig. O.3.: Same subject, but the IMU signal amplitude changes within the same set.

Figure O.4 shows the IMU signals of a subject where the signal signature and amplitude of the performed squat training changes within the same set.

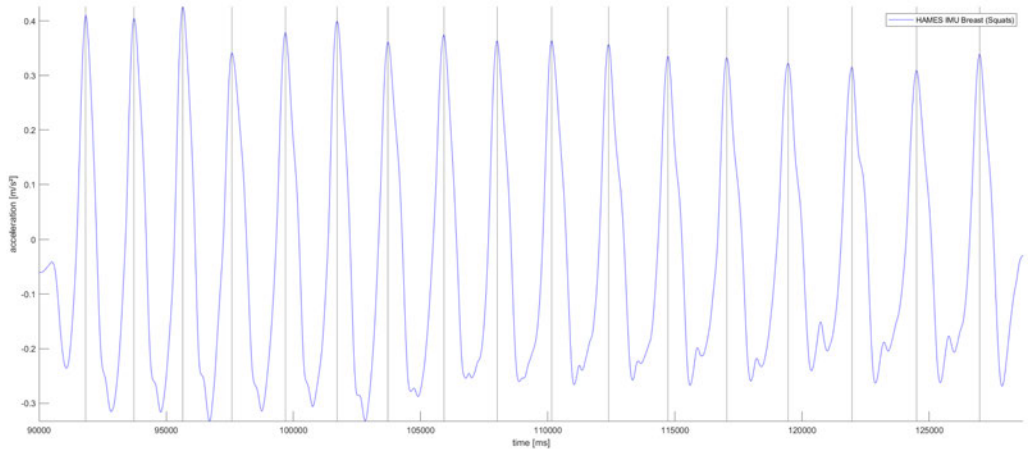


Fig. O.4.: Same subject, but the IMU signal signature and amplitude changes within the same set.

ROC and PR Curves

P

ROC curves plot the true positive rate $\frac{TP}{TP+FN}$ against the false positive rate $\frac{FP}{FP+TN}$. Determining the optimal decision threshold for a model can be challenging. A common approach is to evaluate thresholds for each class individually (one-vs-all) and analyse their performance using ROC curves [57]. ROC curves are particularly useful when class distributions are balanced and false positives and false negatives have similar consequences. They can provide a comprehensive overview of a model's performance across different decision thresholds and facilitate the comparison of classification models by illustrating how well each model discriminates between classes, regardless of class weighting [149, 12, 259].

ROC curves, traditionally used for binary classification, can be adapted for multi-class problems through techniques like one-vs-rest, one-vs-one, micro-averaging, or macro-averaging. However, due to their limitations in multi-class scenarios, alternative visualisation methods (e.g., confusion matrices) and metrics (e.g., F_1 score) often provide more comprehensive insights [2].

In contrast, PR graph plots precision $\frac{TP}{TP+FP}$ on the y-axis and recall $\frac{TP}{TP+FN}$ on the x-axis. Ideally, both high precision and high recall should be achieved, although there is often a trade-off between the two. PR curves are particularly useful for evaluating models in information retrieval scenarios, such as searching a document pool for relevance to a query. They are also preferable for imbalanced data sets, as they focus specifically on the performance of the model with respect to the positive class. In addition, PR curves provide valuable insight into a model's ability to accurately classify positive instances [84, 259].

For both ROC and PR curves, a larger *area under the curve* (AUC) indicates better performance (see Figure P.1). In addition, several measures can be used to summarise the ROC and PR curves into a single comparable metric. For example, the *Equal Error Rate* (EER) is the point on the PR curve where precision equals recall; a lower EER indicates better model performance. Another useful metric can be the *Average Precision* (AP) [57].

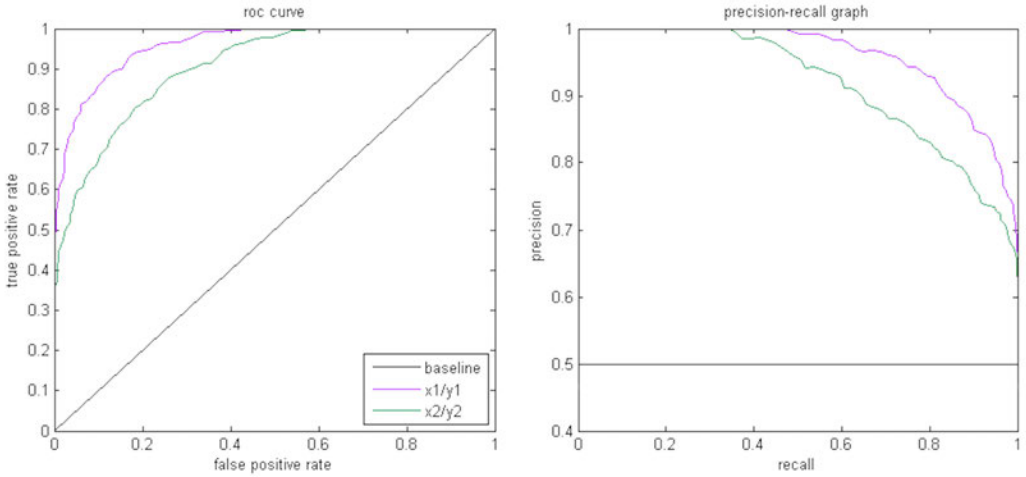


Fig. P.1.: A comparison of ROC and PR curves with two models: x1/y1 and x2/y2. The baseline represents a random classifier.

Data Augmentation Taxonomy

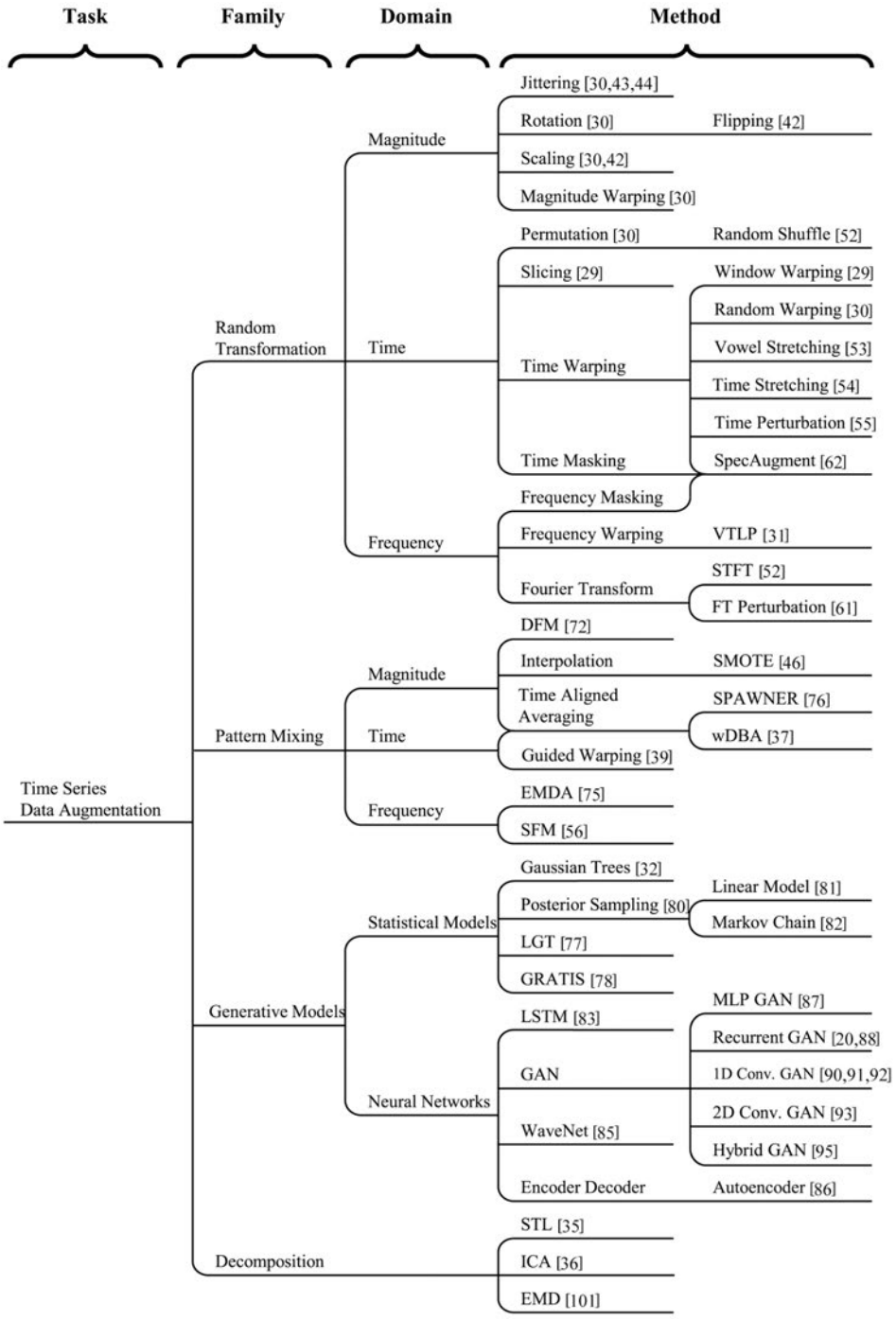


Fig. Q.1.: Taxonomy of time series data augmentation by Iwana and Uchida [166].

R

Forward and Backward Feature Selection

Figure R.1 and Figure R.2 show forward and backward feature selection.

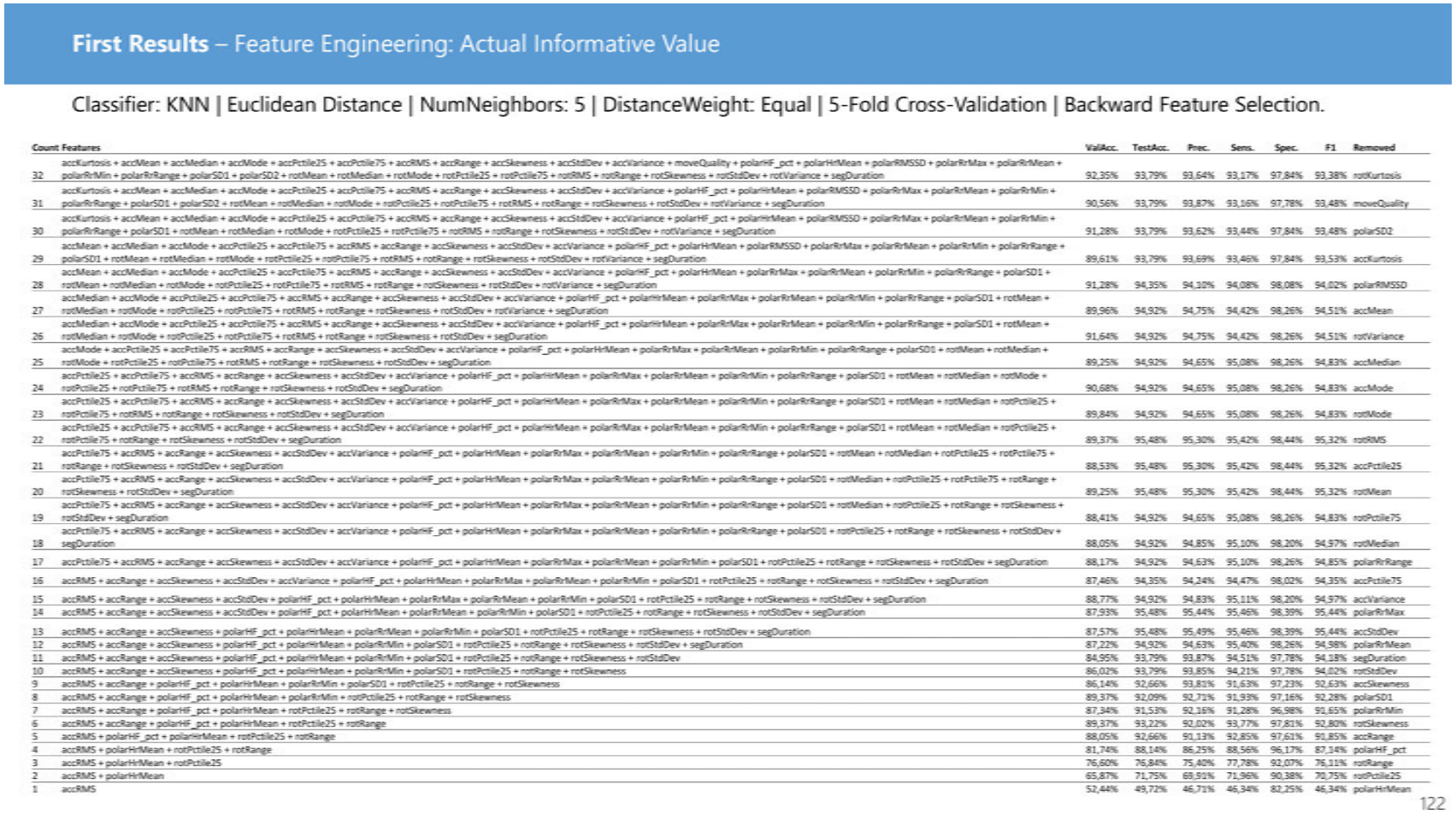
First Results – Feature Engineering: Actual Informative Value

Classifier: KNN | Euclidean Distance | NumNeighbors: 5 | DistanceWeight: Equal | 5-Fold Cross-Validation | Forward Selection.

Count	Features	ValAcc.	TestAcc.	Prec.	Sens.	Spec.	F1	Added
2	accKurtosis + accRange	65,70%	64,97%	66,17%	65,07%	87,80%	65,27%	accRange
3	accKurtosis + accMean + accRange	81,95%	81,92%	82,41%	82,56%	93,51%	82,39%	accMean
4	accKurtosis + accMean + accRange + rotMean	87,14%	88,70%	87,99%	89,23%	96,12%	88,56%	rotMean
5	accKurtosis + accMean + accRange + moveQuality + rotMean	88,73%	89,83%	88,77%	90,23%	96,60%	89,41%	moveQuality
6	accKurtosis + accMean + accRange + moveQuality + rotMean + rotVariance	90,43%	90,40%	89,85%	90,84%	96,77%	90,26%	rotVariance
7	accKurtosis + accMean + accRange + moveQuality + rotMean + rotRMS + rotVariance	90,53%	91,53%	90,58%	91,82%	97,15%	91,16%	rotRMS
8	accKurtosis + accMean + accMedian + accRange + moveQuality + rotMean + rotRMS + rotVariance	91,72%	92,09%	91,56%	92,15%	97,41%	91,75%	accMedian
9	accKurtosis + accMean + accMedian + accRange + moveQuality + rotMean + rotMode + rotRMS + rotVariance	91,23%	94,35%	94,03%	94,11%	98,08%	94,04%	rotMode
10	accKurtosis + accMean + accMedian + accRange + moveQuality + rotMean + rotMedian + rotMode + rotRMS + rotVariance	91,82%	94,92%	94,44%	94,74%	98,32%	94,54%	rotMedian
11	accKurtosis + accMean + accMedian + accRange + accStdDev + moveQuality + rotMean + rotMedian + rotMode + rotRMS + rotVariance	91,23%	94,92%	94,44%	94,74%	98,32%	94,54%	accStdDev
12	accKurtosis + accMean + accMedian + accRange + accStdDev + accVariance + moveQuality + rotMean + rotMedian + rotMode + rotRMS + rotVariance	92,22%	94,92%	94,44%	94,74%	98,32%	94,54%	accVariance
13	accKurtosis + accMean + accMedian + accPctile25 + accRange + accStdDev + accVariance + moveQuality + rotMean + rotMedian + rotMode + rotRMS + rotVariance	92,42%	94,92%	94,44%	94,74%	98,32%	94,54%	accPctile25
14	accKurtosis + accMean + accMedian + accPctile25 + accRange + accStdDev + accVariance + moveQuality + rotMean + rotMedian + rotMode + rotPctile75 + rotRMS + rotVariance	92,12%	94,35%	93,94%	94,40%	98,14%	94,08%	rotPctile75
30	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + moveQuality + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + polarSD2 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotVariance + segDuration	90,64%	93,79%	93,46%	94,41%	97,90%	93,89%	polarSD2
31	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + moveQuality + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + polarSD2 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotVariance + segDuration	91,04%	92,66%	92,49%	92,49%	97,48%	92,45%	polarHF_pct
32	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + moveQuality + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + polarSD2 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	91,04%	93,79%	93,64%	93,17%	97,84%	93,38%	rotStdDev
33	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + moveQuality + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + polarSD2 + rotKurtosis + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	90,80%	92,09%	91,91%	91,55%	97,29%	91,65%	rotKurtosis

27

Fig. R.1.: Forward feature selection with preliminary collected data.



Count	Features	ValAcc.	TestAcc.	Prec.	Sens.	Spec.	F1	Removed
32	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + moveQuality + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + polarSD2 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	92,35%	93,79%	93,64%	93,17%	97,84%	93,38%	rotKurtosis
31	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + polarSD2 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	90,56%	93,79%	93,87%	93,16%	97,78%	93,48%	moveQuality
30	accKurtosis + accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	91,28%	93,79%	93,62%	93,44%	97,84%	93,48%	polarSD2
29	accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRMSSD + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	89,61%	93,79%	93,69%	93,46%	97,84%	93,53%	accKurtosis
28	accMean + accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	91,28%	94,35%	94,10%	94,08%	98,08%	94,02%	polarRMSSD
27	accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + rotVariance + segDuration	89,96%	94,92%	94,75%	94,42%	98,26%	94,51%	accMean
26	accMedian + accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + segDuration	91,64%	94,92%	94,75%	94,42%	98,26%	94,51%	rotVariance
25	accMode + accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + segDuration	89,25%	94,92%	94,65%	95,08%	98,26%	94,83%	accMedian
24	accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotMode + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + segDuration	90,68%	94,92%	94,65%	95,08%	98,26%	94,83%	accMode
23	accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotPctile25 + rotPctile75 + rotRMS + rotRange + rotSkewness + rotStdDev + segDuration	89,84%	94,92%	94,65%	95,08%	98,26%	94,83%	rotMode
22	accPctile25 + accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotPctile25 + rotPctile75 + rotRange + rotSkewness + rotStdDev + segDuration	89,37%	95,48%	95,30%	95,42%	98,44%	95,32%	rotRMS
21	accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMean + rotMedian + rotPctile25 + rotPctile75 + rotRange + rotSkewness + rotStdDev + segDuration	88,53%	95,48%	95,30%	95,42%	98,44%	95,32%	accPctile25
20	accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMedian + rotPctile25 + rotPctile75 + rotRange + rotSkewness + rotStdDev + segDuration	89,25%	95,48%	95,30%	95,42%	98,44%	95,32%	rotMean
19	accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotMedian + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	88,41%	94,92%	94,65%	95,08%	98,26%	94,83%	rotPctile75
18	accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarRrRange + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	88,05%	94,92%	94,85%	95,10%	98,20%	94,97%	rotMedian
17	accPctile75 + accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	88,17%	94,92%	94,63%	95,10%	98,26%	94,85%	polarRrRange
16	accRMS + accRange + accSkewness + accStdDev + accVariance + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	87,46%	94,35%	94,24%	94,47%	98,02%	94,35%	accPctile75
15	accRMS + accRange + accSkewness + accStdDev + polarHF_pct + polarHrMean + polarRrMax + polarRrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	88,77%	94,92%	94,83%	95,11%	98,20%	94,97%	accVariance
14	accRMS + accRange + accSkewness + accStdDev + polarHF_pct + polarHrMean + polarRrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	87,93%	95,48%	95,44%	95,46%	98,39%	95,44%	polarRrMax
13	accRMS + accRange + accSkewness + polarHF_pct + polarHrMean + polarRrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	87,57%	95,48%	95,49%	95,46%	98,39%	95,44%	accStdDev
12	accRMS + accRange + accSkewness + polarHF_pct + polarHrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev + segDuration	87,22%	94,92%	94,63%	95,40%	98,26%	94,98%	polarRrMean
11	accRMS + accRange + accSkewness + polarHF_pct + polarHrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness + rotStdDev	84,95%	93,79%	93,87%	94,51%	97,78%	94,18%	segDuration
10	accRMS + accRange + accSkewness + polarHF_pct + polarHrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness	86,02%	93,79%	93,85%	94,21%	97,78%	94,02%	rotStdDev
9	accRMS + accRange + polarHF_pct + polarHrMean + polarRrMin + polarSD1 + rotPctile25 + rotRange + rotSkewness	86,14%	92,66%	93,81%	91,63%	97,23%	92,63%	accSkewness
8	accRMS + accRange + polarHF_pct + polarHrMean + polarRrMin + rotPctile25 + rotRange + rotSkewness	89,37%	92,09%	92,71%	91,93%	97,16%	92,28%	polarSD1
7	accRMS + accRange + polarHF_pct + polarHrMean + rotPctile25 + rotRange + rotSkewness	87,34%	91,53%	92,16%	91,28%	96,98%	91,65%	polarRrMin
6	accRMS + accRange + polarHF_pct + polarHrMean + rotPctile25 + rotRange	89,37%	93,22%	92,02%	93,77%	97,81%	92,80%	rotSkewness
5	accRMS + polarHF_pct + polarHrMean + rotPctile25 + rotRange	88,05%	92,66%	91,13%	92,85%	97,61%	91,85%	accRange
4	accRMS + polarHrMean + rotPctile25 + rotRange	81,74%	88,14%	86,25%	88,56%	96,17%	87,14%	polarHF_pct
3	accRMS + polarHrMean + rotPctile25	76,60%	76,84%	75,40%	77,78%	92,07%	76,11%	rotRange
2	accRMS + polarHrMean	65,87%	71,75%	69,91%	71,96%	90,38%	70,75%	rotPctile25
1	accRMS	52,44%	49,72%	46,71%	46,34%	82,25%	46,34%	polarHrMean

122

Fig. R.2.: Backward feature selection with preliminary collected data.

Additional Results

S

Figure S.1 shows a box-and-whiskers diagram with accuracy results for k-NN models trained with the same settings but an incremental number of subjects as training set.

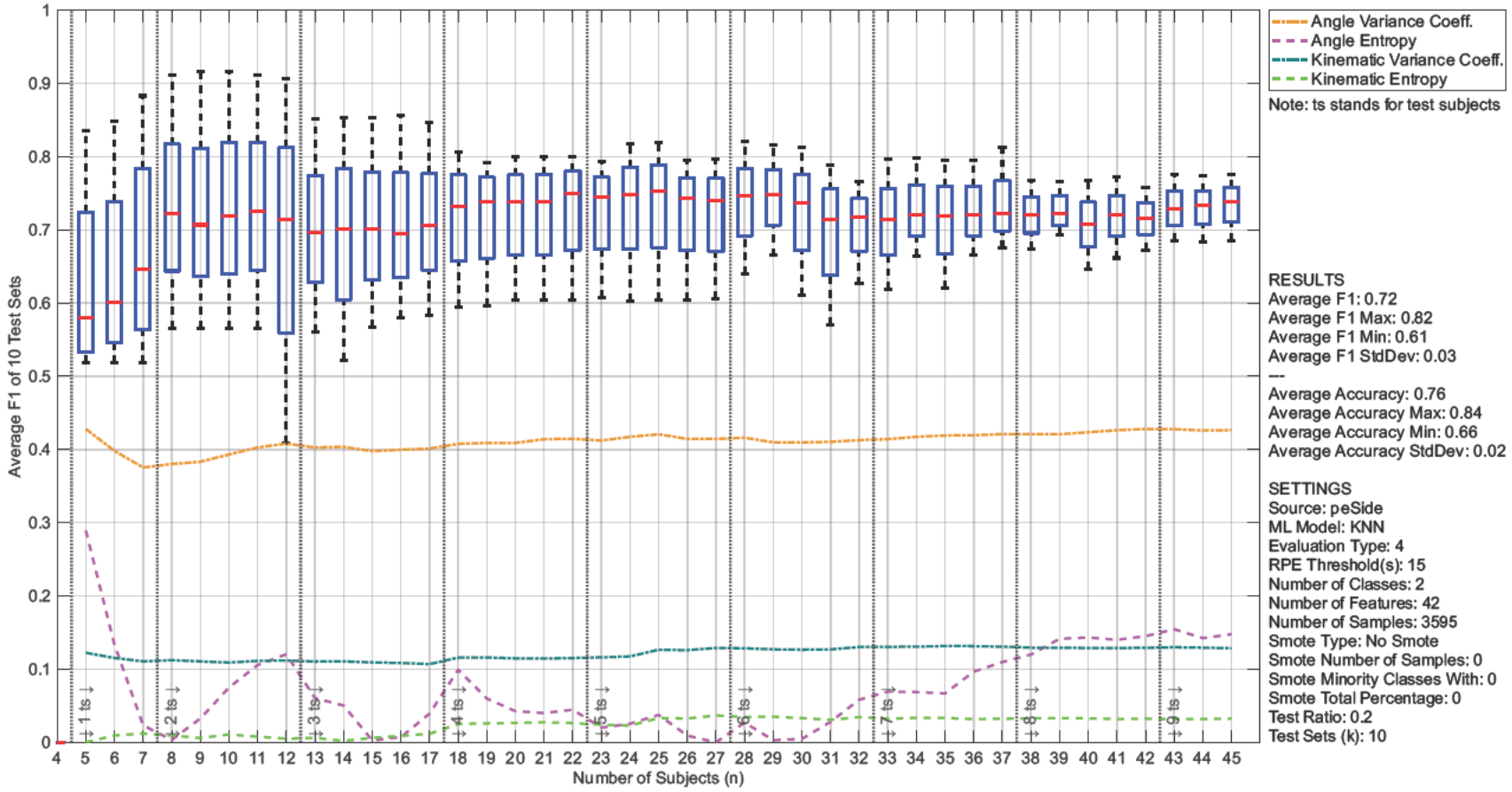


Fig. S.1.: Average accuracy results for an incremental number of subjects (n).