On efficient and precise classification of one-dimensional biomedical ultrasonic signals

Thesis approved by the Department of Computer Science University of Kaiserslautern-Landau
for the award of the Doctoral Degree Doctor of Engineering (Dr.-Ing.)

to

Lukas Brausch

Date of Defense:	27th September, 2023
Dean:	Prof. Dr. Christoph Garth
Reviewer:	Prof. Dr. Paul Lukowicz
Reviewer:	Prof. Dr. Oliver Amft

DE-386

Abstract

In one-dimensional (1-D) Ultrasound (US) measurements, signals are acquired that form the basis of more sophisticated two-dimensional (2-D) or three-dimensional (3-D) US imaging. These 1-D signals contain a lot of raw information about the US wave propagation and interaction with the medium that is only processed in parts during image generation. While image representations are easy to interpret for humans, the analysis of US wave signals is hard to perform without applying algorithms to extract desired features.

This work investigates reliable and fast 1-D US signal classifications to distinguish between different stages or states in biomedical US scenarios and shows how the new field of Machine Learning (ML) on raw US wave data provides advantages and different applications. To achieve good results, the input signals are treated as time series, which requires the deployment of comparatively complex Time Series Classification (TSC) algorithms.

The literature shows that a lot of research efforts have previously only tackled the classification and segmentation of US Brightness mode (B-Mode) images, while neglecting approaches to classify 1-D signals to a large extent. This research contributes by developing, deploying and evaluating classification approaches for three distinct biomedical US classification tasks and finds that respective signal classifications for different scenarios are possible with varying degrees of accuracies. It entails the comparison of several combinations of data types (e.g. temporal, spectral and statistical features or raw signals), ML models and pre-processing steps to provide a strong foundation for robust, binary classifications of 1-D US signals for scenarios based on low-cost wearable, mobile and stationary devices. This research addresses scientific questions not answered before by informing on detailed descriptions of beneficial domain specific knowledge (domain specific knowledge (DSK)), achieved accuracies and times needed for training and evaluation of the examined ML models.

The resulting ML pipelines includes solutions based on data acquired from custom experimental setups or clinical trials. Possible real-world applications might include muscle contraction trackers, muscle fatigue detectors, epiphyseal radius bone closure detectors or devices providing information about advanced liver disease stages. Automated machine-assisted classifications requiring as little DSK as possible from the end user enable application scenarios ranging from fitness or rehabilitation trackers as consumer devices to solutions providing diagnostic support without requiring extensive knowledge from professional medical practitioners. For example, decision support systems for bone age assessments in clinical use or liver health assessment systems for gastroenterologists.

This work shows that reliable, robust and fast classifications based on 1-D US signals are possible with high degrees of accuracies depending on the examined scenario with achieved F_1-scores ranging from $\approx 70\%$ to $\approx 87\%$. These results prove that real-life applications for recreational purposes are already possible and that critical applications for clinical use are highly likely to be achieved once the presented approaches are further optimized in the future.

Acknowledgements

I would like to thank Prof. Dr. Paul Lukowicz from the University of Kaiserslautern and Prof. Dr. Heiko Zimmermann from the Fraunhofer Institute for Biomedical Engineering (IBMT) for giving me the opportunity to work in the exciting field of one-dimensional ultrasound signal classifications. In addition to that, I would also like to thank Steffen Tretbar from the Fraunhofer Institute for Biomedical Engineering (IBMT) for his support and encouragement.

Furthermore, I would also like to thank Prof. Dr. med. Stefan Zeuzem and his team from the University Hospital Frankfurt and the University of Saarland for providing valuable medical data and expertise and partially funding my research.

I would also like to thank Dr. Holger Hewener from Fraunhofer IBMT for his careful supervision, interesting discussions, crucial ideas and helpful feedback. In addition to that I'm also very thankful to all people from Fraunhofer IBMT who volunteered to participate in the experiments described in this work. In particular, I would like to thank all people from the ultrasound department of Fraunhofer IBMT for fruitful discussions and help with setting up the used equipment.

Dr. Axel Schwindling and Dr. Jorge Oevermann kindly provided helpful feedback for the written content of my work, for which I am very grateful.

Finally, I would also like to express my sincere gratitude to my parents Kurt Josef Brausch, Marita Barbara Grün-Brausch, my wife Atuganile Steven Tweve and my son Malte Alexander Brausch for supporting me.

31st October, 2023, Lukas Brausch

Contents

List of Figures

To Malte Alexander Brausch (* 2018). May you live long and prosper.

Chapter 1

Introduction

1.1. Contributions

This work investigates classification strategies to distinguish between different biomedical states or disease stages based on Ultrasound (US) data. A lot of research exists focusing on the classification or segmentation of two-dimensional (2-D) Brightness mode (B-Mode) US images (see Section 2.1.3) but these images require the usage of comparatively expensive and unwieldy US transducers. In contrast to that, one-dimensional (1-D) US signals can be acquired with low-cost single-element US transducers that are feasible for wearable or mobile scenarios. However, research focusing on these signals has been neglected in the past. This work addresses this knowledge gap by providing a comprehensive and thorough evaluation of a variety of different strategies that combine different data types, Machine Learning (ML) models, domain specific knowledge (domain specific knowledge (DSK)) and pre-processing steps. The significance of this research lies in the provision of evidence that robust, fast and accurate binary classifications of 1-D US signals are possible and feasible for wearable or mobile scenarios.

This research applies suitable data science strategies on data acquired from real-world experiments to prove their robustness for the classification of muscle tissue, liver tissue and bone properties. The scenarios examined include US classifications for recreational and fitness applications (see Chapter 3), a dedicated US system that does not implement traditional imaging and analysis (see Chapter 4) and a reporting aid in US typical soft tissue applications (see Chapter 5).

For each scenario, the underlying data was acquired with a custom experimental setup or a clinical study. Healthy subjects participated in the muscle contraction states classification experiments by performing squats to distinguish between contracted and non-contracted muscles and lifted weights chosen according to their subjectively perceived fitness level for as long as possible to induce muscle fatigue. A clinical study with female subjects was conducted at the *Saarland University Medical Center*, including patients from paediatric endocrinology and healthy volunteers to categorize epiphyseal growth plate closures of girls and women based on 1-D US signals acquired with a proprietary mobile device. 1-D US signals acquired with a stationary and commercially available device at the *Frankfurt University Medical Center* from Nonalcoholic fatty liver disease (NAFLD) patients with different fibrosis and steatosis stages served as base for the classification of the respective stages for the identification of hepatic steatosis and fibrosis.

Results from this work show that an average F_1-score of more than 80 % can be achieved for muscle contraction state classifications (distinguishing between relaxed and contracted muscles), muscle fatigue state classifications (distinguishing between relaxed and fatigue muscles), epiphyseal growth plate classifications (distinguishing between open and closed epiphyseal growth plates), liver fibrosis stages (distinguishing between fibrosis stages $\leq F2$ and $F3, F4$) and liver steatosis stages (distinguishing between steatosis stages $S0$ and $S1, S2, S3$) on the acquired data.

These promising findings allow the development of simple devices for muscle contraction or fatigue tracking in fitness or rehabilitation scenarios, US-based finished bone growth detection or low-cost identification of advanced hepatic steatosis or fibrosis. Even though the inherent properties and examined body parts of those biomedical scenarios are fundamentally different from each other, similar data types, pre-processing and DSK can be used to classify signals for wearable and mobile solutions in those distinct fields.

The achieved results in this work are very promising but certain limitations remain. These include technical considerations for wearables devices, challenges concerning data annotations, the need to focus on a suitable subset of ML models and regulatory considerations (see Section 6.4).

1.2. Resulting publications

The work done for this thesis resulted in the following published journal papers, conference papers and databases, sorted chronologically by their respective publication dates.

1.2.1. Journal papers

1. **Lukas Brausch**, Ruth Dirksen, Christoph Risser, Martin Schwab, Carole Stolz, Steffen Tretbar, Tilman Rohrer and Holger Hewener. 2022. *Classification of Distal Growth Plate Ossification States of the Radius Bone Using a Dedicated Ultrasound Device and Machine Learning Techniques for Bone Age Assessments.* In 2022 Applied Sciences 12.7,
doi:10.3390/app12073361. [1]

2. **Lukas Brausch**, Holger Hewener and Paul Lukowicz. 2022. *Classifying Muscle States with One-Dimensional Radio-Frequency Signals from Single Element Ultrasound Transducers.* In 2022 Sensors 22.7, doi:10.3390/s22072789. [2]

1.2.2. Conference papers

1. **Lukas Brausch**, Holger Hewener, and Paul Lukowicz. 2019. *Towards a wearable low-cost ultrasound device for classification of muscle activity and muscle fatigue.* In Proceedings of the 23rd International Symposium on Wearable Computers (ISWC '19). Association for Computing Machinery, New York, USA, 20-22. doi:10.1145/3341163.3347749. [3]

2. Holger Hewener, Christoph Risser, **Lukas Brausch**, Tilman Rohrer and Steffen Tretbar. 2019. *A mobile ultrasound system for majority detection.* IEEE International Ultrasonics Symposium (IUS), Glasgow, United Kingdom, pp. 502-505, doi:10.1109/ULTSYM.2019.8925868. [4]

3. **Lukas Brausch** and Holger Hewener. 2019. *Classifying muscle states with ultrasonic single element transducer data using machine learning strategies.* In Proceedings of the 38rd Meetings on Acoustics. Acoustical Society of America. doi:10.1121/2.0001140. [5]

4. **Lukas Brausch**, Steffen Tretbar and Holger Hewener. 2021. *Identification of advanced hepatic steatosis and fibrosis using ML algorithms on high-frequency ultrasound data in patients with non-alcoholic fatty liver disease.* In 2021 IEEE UFFC Latin America Ultrasonics Symposium (LAUS), pp. 1-4, doi:10.1109/LAUS53676.2021.9639128. [6]

1.2.3. Databases

1. **Lukas Brausch**, Holger Hewener, and Paul Lukowicz. 2019. *21 datasets of raw one-dimensional ultrasound data (A-scans) acquired from the calf muscles of 8 healthy volunteers* [7].

2. **Lukas Brausch**, Holger Hewener, and Paul Lukowicz. 2021. *Datasets of raw one-dimensional ultrasound data (A-scans) acquired from the biceps brachii muscles of 21 healthy volunteers* [8].

3. **Lukas Brausch**, Holger Hewener, and Paul Lukowicz. 2021. *Datasets of raw one-dimensional ultrasound data (A-scans) acquired from the biceps brachii muscles of a single healthy volunteer* [9].

1.3. Outline

This thesis is divided into the following seven chapters, which describe the approaches mentioned above in more detail:

- Chapter 1 **Introduction** introduces and motivates this work.

- Chapter 2 **Background** provides background information and current state of the art technologies used in this work.

- Chapter 3 **Classifying muscle contractions and muscle fatigue** describes strategies to develop, deploy and evaluate ML models for muscle state classifications.
- Chapter 4 **Detection of epiphyseal radius bone closure** describes strategies to develop, deploy and evaluate ML models for epiphyseal radius bone detections.
- Chapter 5 **Identification of hepatic steatosis and fibrosis in patients with non-alcoholic fatty liver disease** describes strategies to develop, deploy and evaluate ML models to identify hepatic steatosis and fibrosis in patients with non-alcoholic fatty liver disease.
- Chapter 6 **Discussion** discusses all results and implications of this work.
- Chapter 7 **Conclusion** concludes this work by summarizing all findings.

Chapter 2

Background

This chapter provides an insight into the theoretical foundation of this work. Section 2.1 explains the physical foundations of ultrasound with a focus on soft tissues. Section 2.2 provides an overview of mobile and wearable technologies in the medical field. Section 2.3 covers the mathematical background needed to understand the methods used in this work, while Section 2.4 covers several available classification models. Section 2.5 provides a comprehensive description of features and how they are used in this work. Section 2.6 and Section 2.7 explains Digital Signal Processing (DSP) and ML, the two major approaches used for classification.

Contents

2.1. Ultrasound

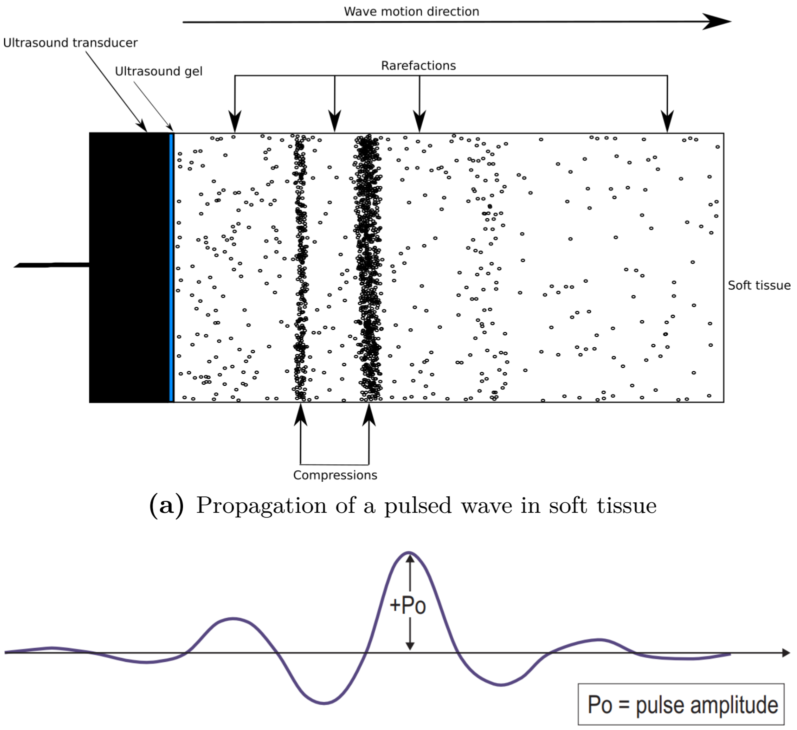
This part introduces various aspects of US and its applications. Section 2.1.1 covers the basic principles and physics. Section 2.1.2 covers Amplitude mode (A-Mode) US and Section 2.1.3 provides an overview of B-Mode US. Section 2.1.4 compares A-Mode and B-Mode US with respect to their respective advantages and disadvantages.

2.1.1. Basic Physics

2.1.1.1. Ultrasound waves

Sound can be defined as oscillation in pressure, stress, particle displacement, particle velocity or the superposition of oscillations propagated in a medium with internal forces (e.g. elastic or viscous). The rate of oscillation is denoted by the term *frequency* and is measured in Hertz (Hz). US waves are sound waves with frequencies above 20 kilohertz (kHz), while sound waves with frequencies lower than 17 Hz are called *infrasound* [10]. Sounds with frequencies below 17 Hz or above 20 kHz are inaudible for human beings but not for some animals, such as bats, dolphins, whales and elephants, which use infrasound or US to navigate and communicate [10, 11]. Sounds with frequencies within this range are audible for human beings. For diagnostic and therapeutic medical applications, typical US frequencies range from 2 to 40 megahertz (MHz) [10].

US waves can be generated by a variety of sources. In medical US, the source is typically one or more *piezoelectric crystals*, which are excited by an alternating voltage. These crystals are able to generate and receive US waves due to the *piezoelectric effect*. Medical US transducers generate short bursts or pulses of vibrations or transmit continuously. The application of an alternating voltage for only a few cycles results in pulsed US, while a continuous application is referred to as continuous US. The former is often used for imaging, while the latter often finds applications in therapy. Figure 2.1 shows the particle displacement of a pulsed wave and its corresponding waveform. The *wavelength* λ is defined as shortest distance between equivalent points on the waveform and the pulse amplitude P_0 stands for the maximal pressure fluctuation. λ can be expressed by the fraction of ultrasound propagation speed c and the frequency f (see Equation 2.1). In soft tissues the US wave induces particle oscillations at its frequency [12].



(a) Propagation of a pulsed wave in soft tissue

(b) Signal representation of a pulsed wave propagation in soft tissue (adapted from Figure 1.3 in [12])

Figure 2.1.: Propagation of a pulsed wave in soft tissue and its corresponding signal representation.

Longitudinal or *compressional* waves move in the direction of wave motion while *transverse* or *shear* waves move perpendicularly to wave motion. Longitudinal waves are much faster than transverse waves and play a pivotal role in classical US imaging and in 1-D US signals used in this work.

Propagation speed of ultrasound waves

The propagation speed of sound in soft tissue is 1540 $\frac{m}{s}$ on average but its exact local value depends on the rigidity and density of the specific medium [13]. This high speed of longitudinal waves enables the acquisition of many US measurements per second. Table 2.1 provides an overview of longitudinal US velocities in different media.

Medium	**Speed** $\frac{m}{s}$
Air	331
Bones	3600
Brain	1530
Fat	1470
Muscles	1568
Water	1492
Soft tissue (average)	1540

Table 2.1.: US speed in different soft tissue media [10].

Equation 2.1 illustrates the relation between the speed of sound c, the wavelength λ and the frequency f [12]:

$$c = f \cdot \lambda. \tag{2.1}$$

As US waves pass from the transducer through a medium, they are subjected to various physical effects such as *diffraction*, *refraction*, *interference*, *scattering*, *attenuation* and *reflection*. Signals echoing back to the transducer are measured and processed. The following part provides an overview of effects contributing to signal measurements in this work affecting the subsequent signal analysis.

Diffraction

A directional change of wave propagation whenever the wave passes through an opening or around a barrier in its path is called diffraction. The US wave spreads according to a certain pattern, which is highly dependent on the shape and size of the source relative to the wavelength of the sound.

Refraction

In contrast to diffraction, refraction occurs when a wave passes a boundary between two media with different sound propagation properties. For a given angle of incidence θ_1, this effect is governed by *Snell's law* [14]:

$$\frac{\sin\theta_1}{\sin\theta_2} = \mu, \tag{2.2}$$

where θ_2 is the angle of refraction and μ the refractive index. Figure 2.2 illustrates the basic principles of reflection and refraction by showing what happens when a wave with incident angle θ_1 of a medium n_1 hits the surface of another medium n_2 with different wave propagation properties. The wave will partly be reflected in an angle equal to the incident angle θ_1 and partly deviate from its original path in an angle θ_2.

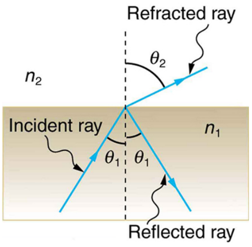


Figure 2.2.: Illustration of the principles of reflection and refraction (adapted from Figure 25.13 in [15]).

Reflection

Figure 2.2 shows that, besides refraction, reflection also occurs when an US wave passes a boundary between two media with different *acoustic impedances* Z. Equation 2.3, applicable for US plane waves, shows that the acoustic impedance of a medium is equal to its *density* ρ times the sound velocity in the medium c [14]:

$$Z = \rho \cdot c. \tag{2.3}$$

The unit of the acoustic impedance is Rayl, where 1 Rayl = 1 $\frac{kg}{m^2 s}$. If an incident US wave is reflected at a flat boundary between two media of acoustic impedances Z_1 and Z_2, the magnitude of the echo amplitude is calculated using Equation 2.4 [12]:

$$A_r = A_i \cdot \frac{Z_1}{Z_2}, \tag{2.4}$$

where A_r is the reflected amplitude and A_i is the incident amplitude. Due to impedance mismatches between different media, the US signals can be severely distorted (e.g. due to the large impedance mismatch between soft tissue and air or soft tissue and bones). This phenomenon is responsible for several limitations of US measurements, such as the inability to penetrate thick bones.

Interference

Interference occurs when several US waves overlap as they pass through the propagating medium. The resultant amplitude of the acoustic pressure at any point is determined by adding the pressure amplitudes from each wave at that point. When the waveforms are in phase, they add constructively and result in an increased amplitude (*constructive interference*). Out of phase, they add destructively and decrease the amplitude (*destructive interference*).

Scattering

As US waves transmit through soft tissue, they interact with small structures whose dimensions are similar to or less than the wavelength λ and whose acoustic impedances exhibit small variations. Whenever smaller structures are hit, some of the energy of the incident waves is scattered in many directions. Figure 2.3 shows the scattering behavior of acoustic waves at rough boundaries between two different media with acoustic impedances Z_1 and Z_2 as well as scattering at inhomogeneities in a medium.

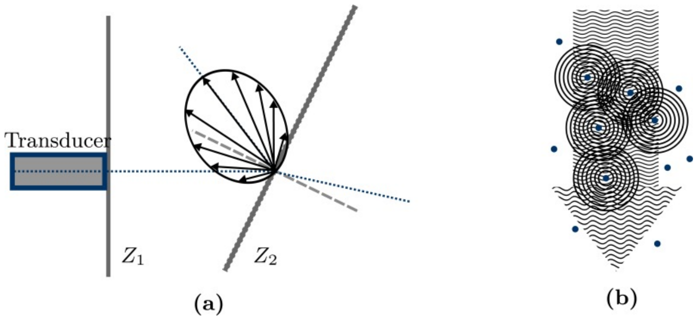


Figure 2.3.: Scattering of sound waves at rough boundaries (a) between two different media with acoustic impedances Z_1 and Z_2 and scattering at inhomogeneities (b) in a medium (taken from Figure 11.3 in [10]).

Speckle patterns in B-Mode US contain spectral information of waves interfering with each other and of the scattering medium that generates the pattern. These speckle patterns are often analyzed for image based classifications.

Attenuation and Absorption

While an US wave penetrates a medium, its intensity J is reduced. This intensity reduction is described by the exponential law of attenuation [10]:

$$J(x) = J_0 \cdot exp(-\mu x), \tag{2.5}$$

where J_0 is the initial intensity. The attenuation coefficient μ stands for the attenuation in decibel (dB) that occurs with each centimeter the sound wave propagates inside a medium. It is influenced by the type of medium and frequency. x represents the currently traveled distance of the US wave. A concept related to attenuation is absorption, which occurs whenever an US wave passes through tissue and its incoming energy is converted into random heat energy, resulting in a reduction of its acoustic pressure. The higher the US frequency, the larger the damping of the amplitude. The tissue specific absorption is also a useful tool for classification (see Section 2.4).

Reflection and Transmission Coefficients

US waves are partly reflected at boundaries of media with differing acoustic impedances. If the acoustic impedance between two media is higher, reflection is greater and transmission is smaller and if the acoustic impedance is smaller, reflection is smaller and transmission is higher. The *reflection coefficient* R is defined as the ratio of reflected wave intensity and transmitted wave intensity. More formally, this can be written as [15]:

$$R = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}, \tag{2.6}$$

where Z_1 and Z_2 stand for the acoustic impedances of the first or second medium, respectively. Since the amount of reflected energy plus the amount of transmitted energy must equal the total amount of incident energy, the transmission coefficient is calculated as follows:

$$T = 1 - \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}. \tag{2.7}$$

2.1.1.2. Ultrasound wave generation and reception

US transducers use the piezoelectric effect to generate and receive US waves. The *piezoelectric effect* describes the conversion of mechanical pressure into an electric charge and vice versa [10, 13]. While the *inverse piezoelectric effect* is used to generate acoustic waves, the *direct piezoelectric effect* is exploited to detect echoes with a measurement system.

In medical US, the transducer transmits US waves into the body. These US waves are the result of vibrations after excitation of the transducer with an electrical source. In most cases, this excitation occurs at the resonance frequency of the transducer. Due to high acoustic impedances, most of the energy of US waves is reflected back at the boundary between soft tissue and air. Usually, ultrasound gel is applied to the skin first to reduce impedance mismatches between different media and increase signal quality. Various physical effects described in Section 2.1.1.1 are responsible for the creation of ultrasound echoes. These acoustic echoes are recorded by the transducer and can be reconstructed as 1-D, 2-D or three-dimensional (3-D) signals, images or videos. The simplest US transducers are single-element piston transducers, which consist only of a single disc shaped piezoelectric element [13].

2.1.2. A-Mode Ultrasound

A-Mode US is the simplest scanning mode, in which *Amplitude scans (A-scans)* of US echoes are recorded. As described above, these A-scans can be obtained with simple single-element US transducers. A-Scans are 1-D signals and contain information about frequency, wavelength, amplitude and wave phase. They can be processed or analyzed in various ways (see Section 2.6) [10]. Unfortunately, A-Scans are often overlain by a variety of different noise sources or physical phenomena (see Section 2.1.1.1), which makes this analysis very challenging.

The intensity of the -data obtained by the US transducer is affected by attenuation and backscattering. These phenomena depend on the type of medium and the US frequency (see paragraph ”Attenuation and Absorption” in Section 2.1.1.1). A frequently used technique to compensate for US attenuation is Time Gain Compensation (TGC) with which certain signal gains are increased to compensate for exponential attenuation and make equally echogenic tissues look the same even if they are located in different tissue depths. To reduce the dependence on user inputs, algorithms for the automatic estimation of attenuation coefficients have been proposed [14, 16].

A collection of geometrical vector envelopes is grouped to an image slice. Typically, logarithmic compression is used to compute the resulting vector, which is then being shown to the user [17]. In abdominal sonography, a broad dynamic range is the most appropriate option for assessing the echotexture of homogeneous soft-tissue structures like liver, pancreas and spleen. Narrow dynamic range is most appropriate for assessing anechoic structures such as aorta and inferior vena cava.

2.1.2.1. Technical applications of A-Mode ultrasound

Nondestructive Testing (NDT) covers a wide range of analytical techniques to inspect, test or evaluate chemical or physical properties of a material, component or system without causing damage. Visual inspection, optical techniques, imaging techniques and the evaluation of electromagnetic fields are popular NDT methods. US testing is a rapidly expanding inspection technique, which relies on the analysis of signals from reflection, transmission and back-scattering of pulsed elastic waves in a material. It uses acoustic waves ranging from 1 kHz to 30 MHz to detect different material flaws and its properties, such as flaw size, crack location, delamination location, fibre waviness, meso-scale ply fibre orientation and layup stacking sequence. A typical US system for NDT consists of a transmitter and receiver circuit, an US transducer and a display device [18].

2.1.2.2. Medical applications of A-Mode ultrasound

Historically, A-Mode US was a popular technique for medical applications. In the 1950s and 1960s, A-Mode US was already used to differentiate breast lesions, to localize liver cysts, to study mediastinal and abdominal masses and to differentiate cysts of the lung and other soft tissue masses [19]. During the 1960s, A-Mode US was considered ”to be a useful procedure in an increasing number of anatomic sites and pathologic conditions” [20] but due to major advances in US B-Mode image quality and developments of 3-D or 4-D US techniques, A-Mode US fell out of favour for many medical applications over the last decades. However, A-Mode US is still used in commercial medical devices in some areas such as ophthalmology, which is a branch of medicine and surgery that deals with the diagnosis and treatment of eye disorders [21]. Other potential or established medical applications for A-Mode US include registration procedures in computer-aided surgeries of the head [22], measurements of fat and muscle thicknesses [23, 24], the quantification of liver fat and diagnosis of NAFLD [25], bone age assessments [26, 27] or cardiovascular screening in general and fetal heart monitoring in particular [28], classification of coronary plaques [29] and the identification of anatomical tissue structures [30]. Additionally, future 3-D US imaging for clinical applications might be enabled by signals from a simple, cheap, single-element transducer in combination with a plastic coding mask by exploiting the signal structure using *compressive sensing* [31]. Even though many applications are available, A-Mode US signals are comparatively unintuitive and hard for humans to interpret due to their 1-D nature.

2.1.3. B-Mode Ultrasound

For the creation of 2-D B-Mode images, transducers with several piezoelectric elements are used. A large variety of transducer types exists, whereby linear, curvilinear and phased array transducers are most common in clinical practice [14]. Figure 2.4 provides a schematic overview of the general layout of a linear array transducer.

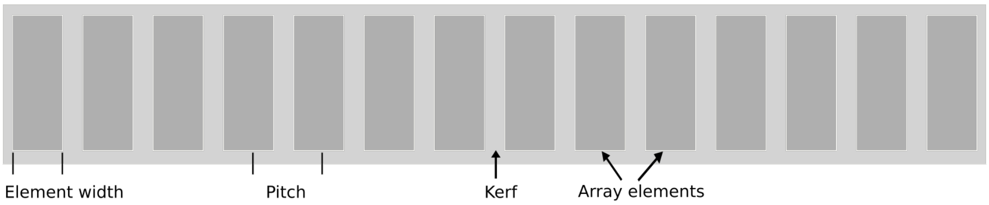


Figure 2.4.: Schematic view of a linear US transducer array.

Each array element can be excited by a custom signal having individual amplitude, phase and waveform properties. Linear US arrays usually contain 64 to 256 elements [13].
Receive beamforming is a spatial reconstruction of the local pressure field amplitudes and the consequent recombination of the received US signals for the purpose of generating images. Transmit beamforming shapes the transmitted beam. In medical US, beamforming is achieved with array transducers. Different beamforming approaches exist and emphasize different features of an US B-Mode image (see Table 2.2) [32].

Imaging feature	Description	Unit
Spatial resolution	Smallest spatial distance for which two close scatterers can be distinguished in the generated image.	mm
Temporal resolution	Time interval between two consecutive images.	Hz
Contrast	Capability to visually delineate different objects in the generated images.	dB
Field of view	Area represented by the obtained images.	cm^2 or cm^3

Table 2.2.: Imaging features important for the choice of beamforming technique [32].

These features are heavily interdependent and each beamforming algorithm is a trade-off emphasizing some features at the cost of others.

2.1.3.1. Technical applications of B-Mode ultrasound

Technical applications of 2-D US imaging are becoming more and more used for industrial NDT [33]. Advantages of using transducer arrays over single-element transducers in NDT are the ability to perform multiple inspections without the need for reconfiguration and the potential for improved sensitivity and coverage. Flexible transducer arrays and high temperature arrays have been developed to allow testing of components with complex geometries and within harsh environments, especially for aerospace and nuclear industries. In addition, air coupled arrays have also shown promising results for nondestructive evaluations of materials. Commonly used B-Mode NDT approaches are the *Total Focusing Method* [34] and *Plane Wave Imaging (PWI)* with the latter being inspired by medical PWI [33].

2.1.3.2. Medical applications of B-Mode ultrasound

B-Mode US is a very common tool for diagnostic medical imaging and is widely available in many healthcare facilities. The major medical fields are [14]:

- **Breast**: Medical imaging of (usually) female breasts.
- **Cardiac**: Medical imaging of the heart.
- **Gynecologic**: Medical imaging of female reproductive organs.
- **Obstetrics**: Medical imaging of fetuses in vivo.
- **Pediatrics**: Pediatric medical imaging.
- **Radiology**: Medical imaging of internal abdominal organs.
- **Sports medicine**: Medical imaging of musculoskeletal structures.
- **Vascular**: Medical imaging of arteries and veins.

2.1.4. Comparison of A-Mode and B-Mode Ultrasound

The analysis of 1-D A-Mode US signals and 2-D B-Mode US images are each afflicted with different advantages and disadvantages. Section 2.1.4.2 and Section 2.1.4.1 aim to discuss these for the respective modes.

2.1.4.1. Advantages and disadvantages of A-Mode ultrasound

Interpretability
A major disadvantage of analyzing 1-D A-Mode signals is that they are, in comparison to 2-D or 3-D visualizations, much less interpretable for humans. Amplitudes of the 1-D echo signal are very unintuitive to analyze even though they contain information about certain physical phenomena or properties of the underlying material or soft tissue. The wave information itself is not interpretable by a human in A-Scans, but humans are good at detecting single significant echoes and reading the depth and thus inferring the depth of large boundary layers just by looking at it. However, in most cases further computer aided analysis is needed.

Classification
In contrast to numerous publications in the active research field of B-Mode US image classification (see Section 2.1.4.2), far fewer publications exist for 1-D US signal classification tasks. The first A-Mode US signal classification approaches solely relied on visual inspections of the signal echoes [19, 20]. However, with the advent of DSP, it became possible to analyze signals in much more detail as described in Section 2.6. In 1983, a publication focusing on breast tissues exploited several preprocessing steps to classify signals

using a *Bayesian decision rule.* These steps included an *envelope extraction* using *Fast Fourier transforms (FFTs)* (see Section 2.3.2.1) and *band-pass filters* before extracting features [35]. In a publication from 1995, A-Mode US was used to characterize intramuscular fat content by relying on *spectral* (see Section 2.3.3.3), *attenuation* (see Section 2.1.1.1), *Kurtosis* (see Section 2.3.3.1) and envelope features as input for computing the *Pearson correlation coefficient (PCC)* (see Section 2.3.3.1) to classify the signals [23]. A publication from 2002 suggests to use intravascular US signals to classify coronary plaques by analyzing spectral features with *autoregressive techniques* and *classification trees* (see Section 2.7.4.3) [29]. Raw 1-D US signals have also been used in the past for the identification of "anatomical tissue structures" [30]. An ophthalmology review publication from 2012 lists the *reflectivity* of A-scans as an appropriate classification feature for a variety of eye pathologies, which can be further processed by DSP algorithms [21].

Wearability
A-Mode US is much better suited than B-Mode US for wearable devices as it does not need highly sophisticated beamforming methods or transducers consisting of several elements. Section 2.2.1 provides more information about currently existing prototypes and research in this field.

Costs
A-Mode US signals can be obtained with single-element transducers, which reduces costs significantly in comparison to B-Mode US. Advanced algorithms to perform beamforming are not required either, which furthermore decreases costs for software engineering.

2.1.4.2. Advantages and disadvantages of B-Mode ultrasound

Interpretability
A major advantage of relying on B-Mode images for diagnosis is that 2-D images are much easier for humans to interpret than 1-D US signals as the former are being optimized for emphasis of visually perceptible structures. This interpretability has sparked a lot of research efforts in the past and made this US mode the most common in clinical usage.

Processing
The classification of medical images in general and B-Mode US images in particular has been addressed by numerous publications and is a very active field of research. Nowadays, the acquisition and storage of large quantities of 2-D images has become comparatively easy and cheap, facilitating 2-D classification tasks. Technological developments and algorithmic advances, such as the introduction of Convolutional Neural Networks (CNNs), revolutionized the field and lead to many applications exploiting massive data collections (i.e. "big data"). In the mid-1970s Kossoff et al. already described that soft tissues can be differentiated by grey scale echography

[36]. Before the advent of ML, traditional *Computer-Aided Diagnosis (CAD)* systems consisting of image preprocessing, image segmentation, feature extraction/feature selection and classification steps as well as general digital image processing algorithms were considered state-of-the-art to *classify* or *segment* medical images [37]. These systems mainly relied on the automated classification of features extracted from texture, morphology or the backscattered echo. Additionally, descriptive features based on examinations of experienced clinicians were also used. Over the course of the last decades, traditional image processing methods have been mostly replaced or enriched by ML methods for B-Mode US images (see Section 2.7) [38, 39]. The initial focus of ML based approaches lay on the classification of extracted features with *linear classifiers*, *Bayesian classifiers*, *Support Vector Machines* (see Section 2.7.4.5), *Decision Trees* (see Section 2.7.4.3), *Artificial Neural Networks (ANNs)* (see Section 2.7.4.6) or methods exploiting ensembles such as *AdaBoost* [40]. However, the introduction of the Convolutional Neural Network (CNN) *AlexNet* [41] shifted the research focus towards Deep Learning (DL) methods, that do not depend on feature extraction as a post-processing step. A review paper published in 2018 mentions 56 publications applying various ML approaches on B-Mode US images of different organs and states that “within the past few years, [DL] approaches have been shown to significantly improve performance when compared with classifiers operating on handcrafted features” [38]. A review paper from 2019 thoroughly discusses a wide range of publications applying DL methods on medical US images for traditional diagnosis tasks including *classification*, *segmentation*, *detection*, *registration*, *biometric measurements*, *quality assessment*, *image-guided interventions* and *therapy*. These methods have been widely applied to different anatomical structures [39]. A publication from 2020 even asserts that “empowered by deep learning, next-generation ultrasound imaging may become a much stronger modality with devices that continuously learn to provide better images and clinical insight, leading to improved and more widely accessible diagnostics through cost-effective, highly-portable and intelligent imaging” [42].

Wearability

Comparatively bulky and expensive acquisition equipment is needed to create B-Mode US images. This goes hand in hand with higher computational costs for the creation of 2-D images. Higher computational costs leads to a higher power consumption, which makes B-Mode US less feasible for wearable solutions as it also reduces battery capacities much faster.

Costs

B-Mode US is always more costly than A-Mode US due to more sophisticated equipment required for the former. For example, a recent study from the USA found that high-resolution B-mode US with proprietary software capable of excluding embedded structures for the measurement of subcutaneous fat thickness “typically exceed[s] US$30,000. Proprietary software and a [2-D] training course are recommended at an additional US$4,000 and US$1,100, respectively” [43].

2.1.4.3. A-Mode ultrasound vs. B-Mode ultrasound trade off

Table 2.3 summarizes all advantages and disadvantages of A-Mode and B-Mode US classifications.

Ultrasound Mode	Power consumption	Equipment bulkiness	Equipment costs	Computational complexity	Interpretability for humans	Availability of classification methods
A-Mode	Low	Low	Low	Low	Low	Low
B-Mode	High	High	High	High	High	High

Table 2.3.: Summary of advantages and disadvantages of A-Mode and B-Mode US.

This work focuses on the processing and classification of 1-D US signals that can be acquired with low-cost, mobile and wearable equipment in contrast to images that are much easier for humans to interpret but rely on more complex and expensive equipment. To illustrate the trade off between A-Mode and B-Mode US, Figure 2.5 shows sample beamformed 2-D data of a gastrocnemius muscle obtained with a linear transducer with 64 elements. For comparison, Figure 2.6a and Figure 2.6b show sample A-scans of a relaxed and fatigue muscle respectively. A single-element transducer from *Panametrics* was used to acquire these signals.

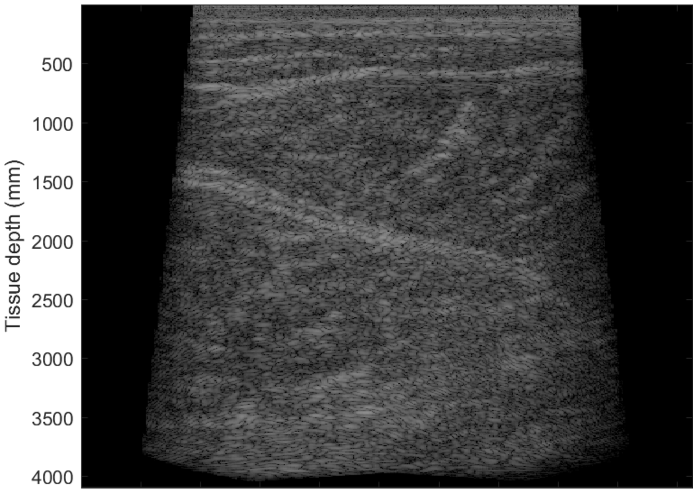
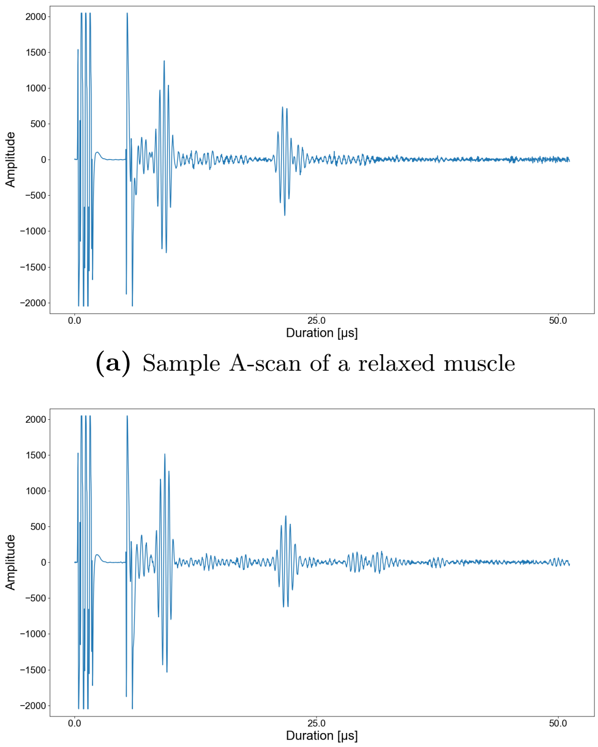


Figure 2.5.: Sample beamformed 2-D data of a gastrocnemius muscle.

Obviously, Figure 2.5 is much more intuitive for humans to interpret. The boundaries of the muscle are clearly visible. Figure 2.6a and Figure 2.6b, however, are much less intuitive. The only information directly retrievable for humans from those A-scans are the amplitude values for any given A-Scan index (i.e. depth). Hence, it is not possible to easily discriminate between two A-scans and categorize them.



(a) Sample A-scan of a relaxed muscle

(b) Sample A-scan of a fatigue muscle

Figure 2.6.: Two sample A-scans of a relaxed (top) and fatigue (bottom) muscle.

2.2. Wearable and mobile medical technology

Wearables are devices that can be worn directly on the body or embedded in fabric, while *mobile devices* are computers that are small enough to be hold and operated in the hand. The construction, evaluation and implementation of such devices are important and heavily researched fields. Wearables and mobile devices are an integral part of digital health solutions and shape them in multiple ways. Digital health refers to the convergence of various information technologies to maximize healthcare effectiveness, including data management (*eHealth*), mobile devices and apps (*mHealth*, *wireless health*) and remote patient management (*connected health*, *telemedicine*). Even though wearable computing is beginning to assume a prominent role in digital health, distributed information sharing and health data integration is still largely a vision for the future as several technical challenges remain [44]. A recent comprehensive review lists several detectable indicators for health monitoring: *body motions* (via *strain sensors* or *pressure sensors*), *body temperature* (via *temperature sensors*), *respiration rate* (via *pressure sensors* or *humidity sensors*), *blood pressure* (via *pressure sensors* or *photodetectors*), *electrophysiological indicators* (via *Electrocardiography (ECG)*, *Electromyography (EMG)* or *Electroencephalography (EEG)*), *metabolites* (via *biosensors*, *Ion sensors* or *pH sensors*), *diseases biomarkers* (via *biosensors*) and *breath analysis* (via *gas sensors*) [45]. Although wearable health monitoring systems have made great progress in recent decades, this study still finds "enormous challenges in scale, multi-function, systematization and intellectualization" of wearable devices [45]. Section 2.2.1 describes current US based wearable solutions as this work aims to facilitate future mobile and wearable US based solutions.

2.2.1. Ultrasound based wearable solutions

Mobile B-Mode (see Section 2.1.3) US devices, such as the products marketed by *Clarius Mobile Health* [46] or *Butterfly Network* [47] cannot be used easily in a wearable fashion due to large transducer surfaces needed for B-Mode US imaging. Wearable B-Mode US devices have been proposed in the past but have, so far, not resulted in commercially available solutions. Examples of wearable devices include a miniaturized device making use of capacitive micromachined ultrasonic transducers (CMUTs) [48] and a system composed of a circular array of 2-D transducers, integrated in a belt, for monitoring the abdominal region and the liver [49]. Additionally, a device for natural sleep recordings for patients with obstructive sleep apnea [50] and an US imaging assembly for routine monitoring of the intima-media thickness, which is a proven indicator of cardiovascular disease [51], have been proposed.
Several portable and wearable systems using A-Mode US (see Section 2.1.2) have been proposed in the past. More recent examples include an US sensor on a paper substrate capable of characterizing "respiratory behavior" [52] or an ultra thin and stretchable US device capable of continuously capturing blood pressure waveforms from deeply embedded arterial and venous sites [53]. Furthermore, a system combining multiple A-Mode US transducers with a conventional motion tracking system to track the motion of bone segments during dynamic conditions [54] has also been developed. Commercially available portable, but not wearable, A-Mode US devices such as the *BodyMetrix BX2000* exist as well [55]. This device is indicated for the measurement of localized fat layer and muscle thickness and has been evaluated by several studies. A publication from 2016 finds this device to be a "reliable tool" in the whole body fat assessment in adults [56]. A more recent work from 2020 partly confirms the reliability of this device for body fat percentage estimates, while also stating that it is more precise for men than for women and that examiner performance is a source of variability that needs to be taken into account [57]. A wearable single-element ultrasonic sensor "made of double-layer polyvinylidene fluoride piezoelectric polymer films with a simple and low-cost fabrication process" consisting of a transmitter and a receiver has been presented [58] and, more recently, the prototype of a "skin-conformal ultrasonic phased array for the monitoring of haemodynamic signals from tissues up to 14 cm beneath the skin" has been reported. "[This] device allows for active focusing and steering of ultrasound beams over a range of incident angles so as to target regions of interest" and "can be used to monitor Doppler spectra from cardiac tissues, record central blood flow waveforms and estimate cerebral blood supply in real time" [59].

2.2.2. Challenges for wearable ultrasound technologies

All US based solutions need ultrasound gel. Its absence would result in severe impedance mismatches between different media, resulting in reflection instead of transmission for the majority of the US waves. This, in turn, would result in very noisy signals. Additionally, more research effort has to be put in the development of completely wireless systems because most current solutions still have to rely on cables to transmit the signals from the transducer to the electronics. Energy consumption, size and thickness of the US transducers are also important factors that need to be taken into account to achieve a good user experience.

2.3. Mathematical foundations

This section aims to provide an overview of mathematical foundations needed for this work. Section 2.3.1 discusses different aspects concerning time series analysis in general, whereas Section 2.3.2 introduces all mathematical transforms and filters used in various ML methods. Section 2.3.3 provides detailed insights about the computation of a variety of features.

2.3.1. Time series analysis

A time series $X = \{x_0, x_1, ..., x_{n-1}\}$ is a discrete sequence of data points indexed by time, most commonly, at successive equally spaced points. Time series analysis aims to extract meaningful statistics and other characteristics of the data. *Time Series Forecasting* aims at making predictions, while *Time Series Classification (TSC)* aims at classifying time series into distinct categories. The classification of time series is of interest in various fields, such as speech recognition, financial analysis, manufacturing, power systems, electronic health records, human activity recognition, acoustic scene classifications and even cybersecurity [60, 61].

Section 2.4 discusses TSC more thoroughly, while Section 2.5 focuses on several processing steps related to features. *DSK* is very important to extract meaningful features by exploiting knowledge obtained from related research [62]. Chapter 5 shows an example of DSK being integrated into the classification pipeline as the algorithms used for the liver disease stage classification only work on A-scans of a certain soft tissue depth. By truncating the A-scans to a certain depth, the DSK that the human liver has to be located in a restricted area within the body is exploited. In the following, every A-scan A_n is a time series represented as an 1-D vector containing n samples. This allows to deploy TSC algorithms and methods on 1-D US signals. A vast variety of different TSC algorithms exists and choosing algorithms with a good performance with respect to accuracy and speed is often a trade-off requiring compromise (see Section 2.7.4).

2.3.2. Mathematical transforms and filters

A mathematical transform is a function f that maps a set X to itself, i.e. $f : X \rightarrow X$. To perform TSC, it can be advantageous to perform one or several mathematical transforms first before processing the signals further, as different time series representations can simplify time series comparisons and reduce the dimensionality of similarity searches [63]. A large variety of different types of transforms exists but the following sections only restrain themselves to those that have been used for TSC tasks in this work. Section 2.3.2.1, Section 2.3.2.2 and Section 2.3.2.3 provide an overview of the *Fourier Transform*, the *Wavelet Transform* and the *Hilbert Transform* respectively. The *Discrete Fourier transform (DFT)* and the *Discrete Wavelet transform (DWT)* are commonly used methods [63], while the Hilbert transform is especially important in the context of US signal processing [14].

2.3.2.1. Fourier transform

The Fourier transform, named after the French mathematician and physicist Jean-Baptiste Joseph Fourier, decomposes a function depending on space or time into a function depending on spatial or temporal frequency. In particular, the DFT is a form of Fourier analysis that is applicable to a sequence of discrete values. The DFT coefficients F_k of a time series $X = \{x_0, x_1, ..., x_{n-1}\}$ are complex numbers given by the following formula

$$F_k = \sum_{i=0}^{N-1} x_i \cdot e^{\frac{-j2\pi ik}{N}}, \tag{2.8}$$

in which j is the imaginary unit. Its inverse operation is given by the formula

$$x_n = \frac{1}{N} \sum_{i=0}^{N-1} F_k \cdot e^{\frac{j2\pi ik}{N}}. \tag{2.9}$$

One advantage of using DFT for signal processing is that it can leverage the computational complexity of the Fast Fourier transform (FFT), which is $\mathcal{O}(n \log n)$ [63]. Fourier transformed signals are very important for the computation of some time series features (see Section 2.5 for details) and can also serve as input data. Thus, applying the FFT to a signal before further processing it in a ML pipeline can serve as a step to reduce the signal's dimensionality first.

2.3.2.2. Wavelet transform

The Wavelet transform utilizes basis functions called *wavelets* that allow the localization of time series in frequency and space [63]. The Hungarian mathematician Alfréd Haar initiated the development of wavelets by introducing a function known today as Haar wavelet. Different families of wavelets (also called *mother wavelets*) have different trade-offs with respect to compactness and smoothness. See Figure 2.7 for an overview of different discrete and continuous wavelet families.

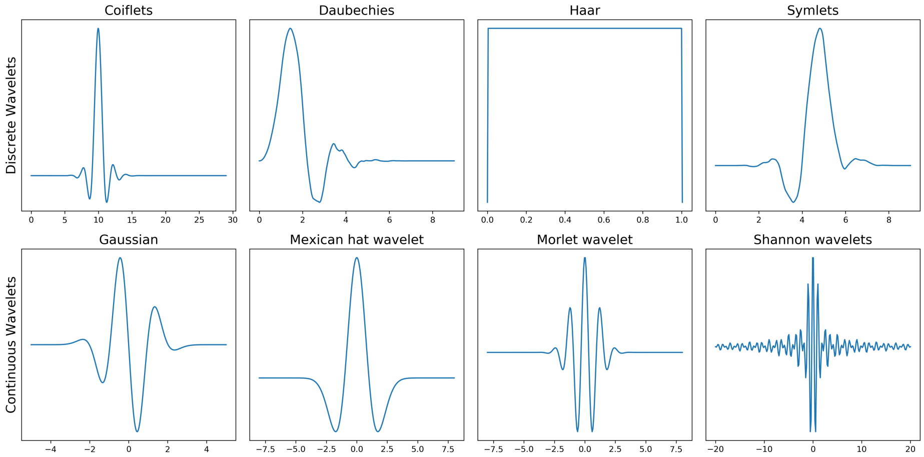


Figure 2.7.: Examples of discrete (top row) and continuous (bottom row) wavelet families.

Each mother wavelet Ψ can have different child wavelets, which can be generated by applying the following formula [63]:

$$\Psi(t)^{s,\tau} = \frac{1}{\sqrt{s}} \cdot \Psi(\frac{t-\tau}{s}), \tag{2.10}$$

where s and τ are the contraction and translation constants “used to slide a window over the time series”. The wavelet transform can analyse time series at different scales, which is “a significant advantage over the Fourier transform, whose basis functions (sines and cosines) do not allow any time series localization in space” [63] (see Section 2.3.2.1). In practice, the DWT is usually applied by multiplying the input signal with the chosen wavelet at different time locations, resulting in a convolution of the signal. Stacking these 1-D convolutions iteratively results in a 2-D *spectrogram*.

2.3.2.3. Hilbert transform

The Hilbert transform, named after the German mathematician David Hilbert, is capable of computing the analytical signal of raw US signals. It takes a function $f(t)$ of a real variable and produces another function of a real variable $g(x)(t)$. The Hilbert transform is defined as follows [64]:

$$g(x) = \frac{1}{\pi} \cdot \int_{-\infty}^{\infty} f(u) \cdot \frac{1}{x-u} du. \tag{2.11}$$

The function $h(x) = \frac{1}{\pi x}$ is called the *convolution kernel* and is singular at $x = 0$. The Hilbert transform is strongly related to the Fourier transform (see Section 2.3.2.1) and can be computed with the following steps [64]:

1. Calculate the Fourier transform of the input signal $x(t)$.
2. Reject the negative frequencies.
3. Calculate the inverse Fourier transform. The real and the imaginary parts resulting from this step are called the *Hilbert transform pair*.

A Hilbert transform pair consists of two functions $f(x)$ and $g(x)$, such that $g(x)$ is the Hilbert transform of $f(x)$ and $-f(x)$ is the Hilbert transform of $g(x)$. The Hilbert transform is a common way to determine the envelope of a signal. A narrow-band signal $s_R(t)$, can be factored as a product of the slow-varying envelope $A(t)$ and fast-varying fine structures $f(t)$:

$$s_R(t) = A(t) \cdot f(t) = A(t) \cdot cos(\phi(t)), \tag{2.12}$$

where $d\phi(t)/dt$ is the instantaneous frequency of the signal [64].

2.3.2.4. Band-pass filter

A *filter* is an operation that produces each sample of the output waveform y as a weighted sum of several samples of the input waveform x as follows [65]:

$$y(t) = \sum_{n=0}^{N} h(n)x(t-n). \tag{2.13}$$

Here, t is the analysis point in time and $h(n)$ is the impulse response. *Band-pass filtering* in particular is used to isolate a selected frequency range of the US transducer *passband*, which is the range of frequencies or wavelengths that can pass through a filter. An ideal band-pass filter would allow through all frequencies within a completely flat passband without amplification or attenuation. Such a filter would completely attenuate all frequencies outside the passband. The *bandwidth* of a filter is the difference between the upper and lower cut-off frequencies. A common band-pass filter is the *Butterworth* filter, named after the British engineer and physicist Stephen Butterworth [14, 65].

Figure 2.8 provides a comparison of the effects of different transformations and filters on A-scans acquired for the muscle fatigue classification scenario in this work (see Chapter 3). It shows a signal belonging to a relaxed muscle state (left column) and a signal belonging to a fatigue muscle state (right column). The following summarizes the depicted signals:

- Row 1: Raw A-scans.
- Row 2: Band-pass filtered A-scans using the Butterworth algorithm.
- Row 3: Raw A-scans transformed with a FFT.
- Row 4: Raw A-scans transformed with a Wavelet transform.
- Row 5: Raw A-scans transformed with a Hilbert transform.

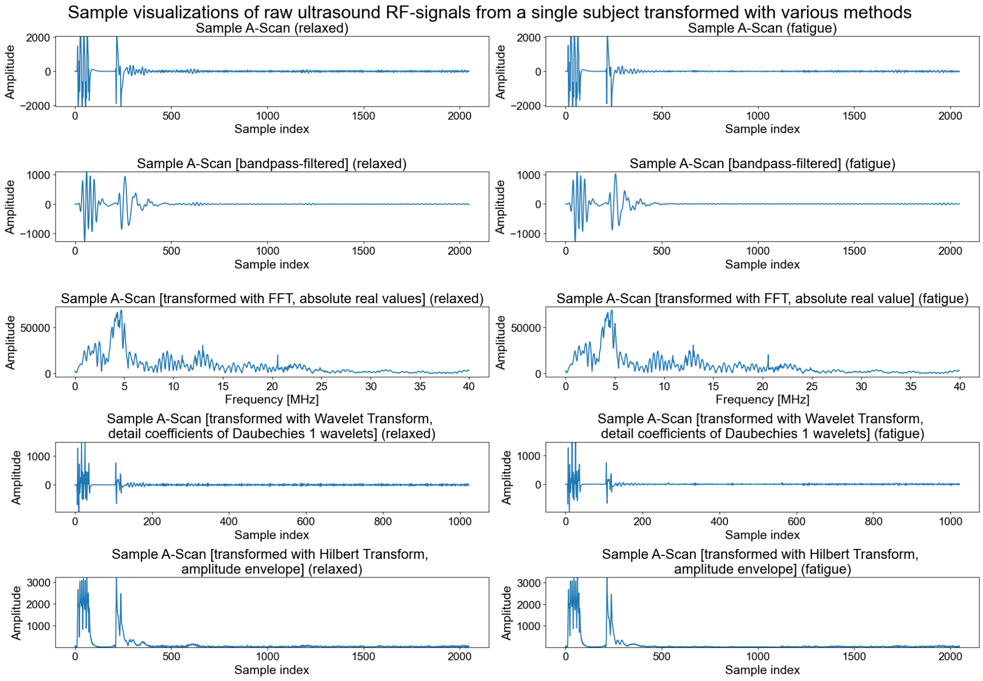


Figure 2.8.: Comparison of the effects of different transformations and filters on A-scans acquired for a muscle fatigue classification scenario.

2.3.3. Computation of features

This section provides a comprehensive overview of the computational details of different features that can be used for DSP or ML models. A feature is a, usually numeric, property or characteristic of a phenomenon. Features can be computed including or excluding DSK. Selecting or computing informative, distinct, discriminating and independent features is a crucial foundation for many effective algorithms in TSC. Section 2.5 discusses the engineering (see Section 2.5.1), extraction (see Section 2.5.2), learning (see Section 2.5.3), selection (see Section 2.5.4) and scaling (see Section 2.5.5) of features in detail. This section restricts itself to a comprehensive explanation of the features used in the TSFEL library [66] as these are extensively used for the models of this work.

2.3.3.1. Statistical features

This section provides the definitions of important concepts used to compute or derive statistical features for the models used in this work.

- **Empirical cumulative distribution function (ECDF)**: The distribution function associated with the empirical measure of a sample [67].
- **Histogram**: An approximate representation of a numerical data distribution dividing the value range into consecutive and

non-overlapping intervals [66, 68]. A histogram gives an estimate of the posterior density of a numerical sample [67].

- **Interquartile range (IQR)**: A measure of statistical dispersion, based on dividing a dataset into four equal quartiles. Given a vector V of length n, the q-th percentile of V is the value $\frac{q}{100}$ of the way from the minimum to the maximum in a sorted copy of V. The values and distances of the two nearest neighbors as well as the interpolation parameter will determine the percentile if the normalized ranking does not match the location of q exactly. The IQR is the third quartile subtracted from the first quartile [66, 68].
- **Kurtosis**: $\frac{\mu_4}{\sigma^4}$, where μ is the fourth central moment and σ is the standard deviation. It describes the "tailedness" (i.e. shape of the probability distribution of a real-valued random variable) [66, 69].
- **Root Mean Square (RMS)**: If n values $x_1, x_2, \cdots, x_n$ are given, the RMS is computed as follows: $x_{RMS} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)}$.
- **Skewness**: A measure of a probability distribution's asymmetry of a real-valued random variable about its mean. The skewness value can be positive, zero, negative or undefined. For normally distributed data, the skewness should be about zero. For unimodal continuous distributions, a skewness value greater than zero means that there is more weight in the right tail of the distribution [66, 69].
- **Standard deviation**: The standard deviation measures the amount of variation in a set of values. A low standard deviation means that the values are close to the mean of the set, while a large standard deviation means that the values are spread out over a wider interval [66, 68].
- **Variance**: The expectation of the squared deviation of a random variable from its mean. It measures how far a set of numbers is spread out from their average value. The variance of a collection of n equally likely values can be written as: $Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, where μ is the mean [66, 68].

2.3.3.2. Temporal features

This section provides the definitions of important concepts used to compute or derive temporal features for the models used in this work.

- **Absolute Energy**: The sum of all squared absolute values of a set of n values $x_1, x_2, \cdots, x_n$ [66, 68].
- **Area Under the Curve (AUC)**: An approximation for the region under the graph of the function $f(x)$ with starting point a and ending point b [66, 68].
- **Autocorrelation**: The correlation of a signal with a delayed copy of itself. This measure can be described as similarity between observations as a time lag function between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its

harmonic frequencies [66, 68].

- **(Spectral) Centroid**: A measure to characterize a spectrum and indicate where the center of mass of the spectrum is located. It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights [66, 68].
- **(Spectral) Entropy**: The spectral entropy treats the signal's normalized power distribution in the frequency domain as a probability distribution and calculates the *Shannon entropy* of it [66, 68].
- **Mean Absolute Difference**: A measure of the average absolute difference of two independent values drawn from any probability distribution [66, 68].
- Positive or negative **turning points** of a signal are points where the function's derivative is zero. A turning point may be either a **local minimum** or **local maximum** [66, 68].
- The **peak to peak distance** of a signal is the absolute value of the difference between its maximum and minimum points [66, 68].
- The **total traveled distance** of a signal is defined as the hypotenuse of a virtual triangle between 2 datapoints [66, 68].
- The **slope** of a signal is computed by fitting a linear equation using a polynomial $p(x) = p_0 + p_1 * x^1 + \cdots + p_k * x^k$ of degree k [66, 68].
- The **total energy** of a signal is computed as follows [66, 68]: $E_x = \sum_n |x_n|^2$.
- **Zero-crossing rate**: The rate at which a signal changes from positive to zero to negative or from negative to zero to positive [66, 68].
- The **neighbourhood peaks** of a signal calculates the number of peaks of at least support n. A peak of support n is defined as a subsequence of the signal in which n neighbour values of a given signal index to the left and to the right are larger than the value at this index [66, 70].

2.3.3.3. Spectral features

This section provides the definitions of important concepts used to compute or derive spectral features for the models used in this work.

- **Fourier series coefficients** c_k are given by:

$$c_k = \int_{-\frac{1}{2}}^{\frac{1}{2}} xe^{-i2\pi kx} dx = i\frac{\cos(\pi k)}{2\pi k},$$

  for $k = 0, \pm 1, \pm 2, \cdots$. The Fourier coefficients are imaginary but the discrete FFT can be used to approximate them [71].
- The **fundamental frequency** δw is defined as the lowest frequency of a periodic waveform [71].
- The **human range energy ratio** is the ratio between energies in the frequency range [0.6–2.5Hz] and the whole energy band [66]. This feature is called human range energy ratio due to its significance in human speech recognition and classification.

- **Linear prediction cepstral coefficients** are Fourier transform coefficients illustrating the logarithmic magnitude spectrum. This feature is commonly applied in the field of speech processing because of its ability to perfectly symbolize speech waveforms and characteristics with a limited set of features [72].
- **Mel Frequency cepstral coefficients** were originally suggested for identifying monosyllabic words in continuously spoken sentences. Their computation attempts to replicate the human hearing system intending to artificially implement the ear's working principle with the assumption that the human ear is a reliable speaker recognizer [72].
- The **power spectrum** $P(\omega)$ of a signal is the distribution of power into frequency components of that signal. It can be computed by using the continuous Hartley transform [71].
- The **power spectrum density bandwidth** B of a signal corresponds to the width of the frequency band in which 95 % of its power is located [66].
- The **spectral decrease** averages the set of slopes between frequency f_k and f_1. It therefore emphasizes the slopes of the lowest frequencies [66, 73].
- The **spectral distance** is the distance of a signal's cumulative sum of the Fourier transform elements to the respective linear regression.
- The **spectral kurtosis** μ_4 is a measure of the spectrum flatness around its mean value. $\mu_4 = 3$ indicates a normal (Gaussian) distribution, $\mu_4 < 3$ a flatter distribution and $\mu_4 > 3$ a peakier distribution [66, 73].
- The **spectral positive turning points** are the number of positive turning points of the magnitude signal of the Fourier transform . See Section 2.3.3.2 for a definition of the turning points.
- The **spectral roll-off** is defined as the frequency $f_c(t_m)$ below which 95 % of the signal energy is contained [66, 73].
- The **spectral roll-on** is defined as the frequency $f_c(t_m)$ below which 5 % of the signal energy is contained [66].
- The **spectral skewness** μ_3 is a measure of the asymmetry of the spectrum around its mean value. $\mu_3 = 0$ indicates a symmetric distribution, $\mu_3 < 0$ represents more energy at frequencies lower than the mean value and $\mu_3 > 0$ means more energy at higher frequencies [66, 73].
- The **spectral slope** is computed using a linear regression over the spectral amplitude values [66, 73].
- The **spectral spread** or spectral standard-deviation μ_2 represents the spread of the spectrum around its mean value [66, 73].
- The **spectral variation** represents the amount of spectrum variation over time [66, 73].
- The **wavelet absolute mean** values represent the absolute mean value of each wavelet scale [66].
- The **wavelet energy** values represent the variation in signal intensity. It is assumed that each amplitude in the signal will demonstrate a

distinctive range of wavelet energy values. Wavelet energy in time line directions is defined as: $E_i^t = \sum_{x=1}^{p}(T_i(x))^2$, where p is the frequency expressed in terms of radians [66, 74].

- The **Shannon wavelet entropy** is calculated by: $E = -\sum_{i=1}^{M} d_i log d_i$, where $d_i = \frac{|W(a_i,t)|}{\sum_{j=1}^{M} W(a_j,t)}$ with $W(a_i,t)$ being the wavelet coefficient at scale a_i [66, 75].
- The **wavelet standard deviation** for scale a_i is defined to be the standard deviation of the wavelet coefficients $W(a_i,t)$ [66].
- The time-dependent **wavelet variance** for scale a_i is defined to be the variance of the wavelet coefficients $W(a_i,t)$ [66, 76].

2.4. Classification

The objective in classification is to assign any input data vector a discrete category, class or group. Most commonly, each input vector can only be assigned to one single class [77]. This work puts an emphasis on the binary classification of 1-D US signals by applying TSC methods. This is not trivial and has been named one of “10 challenging problems in data mining research” in 2006 [78]. Similar statements have been reiterated over the years, with a paper from 2019 even stating that TSC “is a hard problem that is not yet fully understood and numerous attempts have been made in the past to create generic and domain specific classification methods. Because of the diverse domains where time series are present, the research and methods are diverse as well” [62]. The *UCR Time Series Archive* was first introduced in 2002 and remains an important benchmark tool to compare TSC algorithms [79]. Figure 2.9 shows a taxonomy of different TSC methods, including Support Vector Machines (SVMs), Random Forests (RFs), Symbolic Aggregate approXimation (SAX), 1-Nearest Neighbor with Dynamic Time warping (1-NN DTW) and DWT. In this figure, *ED* is an abbreviation for *Euclidean Distance*.

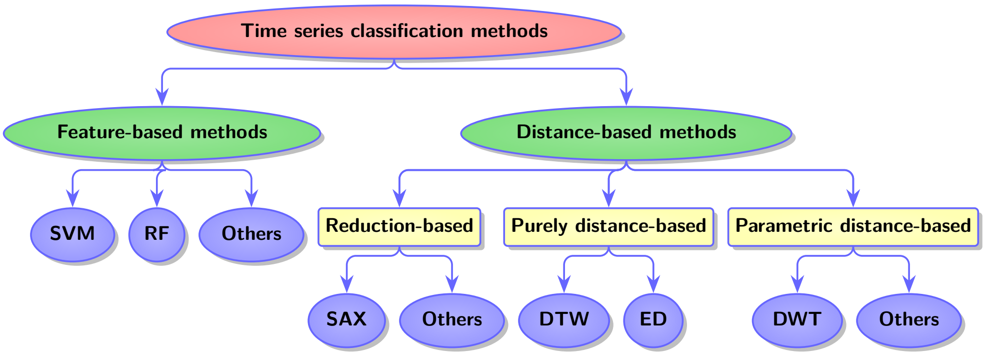


Figure 2.9.: Alternative taxonomy of TSC methods (adapted from [60]).

This taxonomy distinguishes between the two branches feature-based (FB) and distance-based (DB) methods. The former methods perform feature extraction before classification (see Section 2.5.1), while the latter avoid the feature extraction phase and compare signals based on suitable *distances*. Section 2.4.1 provides an overview of FB methods and Section 2.4.2 discusses DB methods. Section 2.4.3 discusses *dictionary-based* approaches, which transform the input signals into representative words. Section 2.4.4 introduces shapelets, which are subsequences of time series that are discriminatory of class membership, while Section 2.4.5 discusses algorithms that derive features from time series intervals. Section 2.4.6 provides an overview of ensemble methods consisting of several classifiers. Another possibility to perform TSCs is to use ANNs, which are a special case as recent methods can avoid the feature extraction phase by relying on an *end-to-end* pipeline even though ANN-based classifiers have traditionally often relied on extracted features. Section 2.4.7 introduces (deep) ANNs directly working on 1-D input signals. Section 2.4.8 discusses ANNs transforming the input signals first to higher dimensional data, e.g. images, before performing classification tasks. Section 2.7.4 discusses the methods used in this work in greater detail.

2.4.1. Feature-based methods

FB methods rely on features (see Section 2.5) to classify signals. The most common FB classification approaches are the k-nearest neighbors algorithm (kNN), SVMs, Relevance Vector Machines (RVMs), Decision Trees (DTs), RFs, Logistic Regression (LR), Gaussian Processes (GPs) and (deep) ANNs [60]. The k-NN algorithm finds major application in a range of warping or editing based distance measures, such as the *1-Nearest neighbor with Dynamic Time warping* algorithm, which is discussed in Section 2.7.4.2. Section 2.7.4.3 introduces DTs, which play a crucial role in Gradient Boosting Machines (GBMs) (see Section 2.7.4.4). Section 2.7.4.6 discusses the ANN methods Multilayer Perceptrons (MLPs) and CNNs, which provide competitive TSC performances [61].

2.4.2. Distance-based methods

Figure 2.10 shows an advanced and extended taxonomy of distance based TSC methods [80].

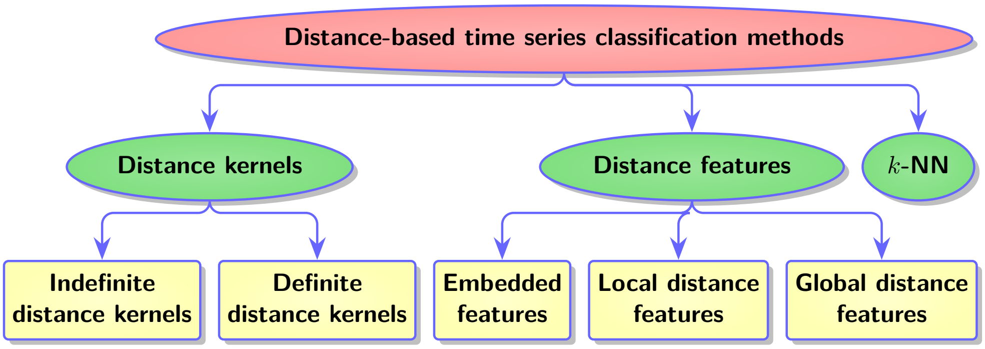


Figure 2.10.: Taxonomy of DB TSC methods (adapted from [80]).

This taxonomy distinguishes between *distance kernels*, *distance features* and *k*-NN. A widely known and deployed distance-based method is 1-NN DTW. Alternative DB models to 1-NN DTW are *weighted DTW*, *Time Warp Edit*, *Move-Split-Merge*, the *Complexity Invariant Distance*, *Derivative DTW* and *Derivative Transform Distance* [81]. Many different publications have shown that algorithms based on DTW distance "[seem] to be particularly difficult to beat" [80]. Even though it has been shown that Support Vector Machine (SVM) based approaches using sophisticated kernels outperform 1-NN DTW, the corresponding experiments have only been performed "with just two metric and one non-metric measures which is not enough to draw strong conclusions" [80]. This work includes the 1-NN DTW algorithm in further analyses, as this algorithm is a popular benchmark and previous work has shown that many proposed alternatives either result in lower or only slightly higher average accuracies [81]. Furthermore, the performance of (1-NN) DTW "can be improved with very little effort" and in many cases "simple improvements can close most or all the [gaps] between DTW and more complex [methods]" [79].

2.4.3. Dictionary-based methods

Dictionary-based TSC methods "approximate and reduce the dimensionality of [time] series by transforming them into representative words, then basing similarity on comparing the distribution of words" [81]. In a comprehensive comparison on the UCR archive, several Dictionary-based methods were thoroughly examined. Of those, Symbolic Aggregate Approximation - Vector Space Model (SAXVSM) and Bag of Patterns (BOP) performed worse than the 1-NN DTW benchmark [81]. The algorithm Bag-of-SFA-Symbols (BOSS) yielded better results than 1-NN DTW and its ensemble version even performed significantly better than other classifiers on several datasets of the UCR classification benchmark in its 2015 introductory publication [81, 82]. Unfortunately, BOSS "has a training complexity quadratic in both the number of training examples and time series length, $\mathcal{O}(n^2 \cdot l^2)$" [83]. The Dictionary-based Word ExtrAction for time SEries cLassification (WEASEL) algorithm has been introduced in 2017 as being "more accurate than the best current non-ensemble algorithms at orders-of magnitude lower classification and training times and it is almost as accurate as ensemble classifiers, whose computational complexity makes them inapplicable even for mid-size [datasets]" [84].

2.4.4. Shapelet-based methods

Classification methods based on *shapelets* "focus on finding short patterns that define a class, but that can appear anywhere in the [time] series. These phase independent patterns are commonly called shapelets" [81]. A publication from 2009 introduced the concept of *time series shapelets* for the first time and stated that shapelet-based algorithms can be interpretable, more accurate and significantly faster than state-of-the-art classifiers [85]. Later work published in 2012 built upon the ideas presented in [85] by extracting the k best shapelets and using them “to transform the data by calculating the distances from a time series to each shapelet” [86]. Both aforementioned approaches evaluate how well a shapelet distinguishes between classes. However, often a shapelet is most useful in distinguishing between members of the class of the time series it was drawn from against all others. Incremental improvements to mitigate these problems of the shapelet transform, specifically for multi-class problems, have been proposed in a 2015 publication that proposes a method simplifying quality assessment calculations, speeding up the execution and increasing the accuracy for multi-class problems [87].

2.4.5. Interval-based methods

Interval-based methods perform classifications based on information contained in various different intervals of a given time series. In 2013, the Time Series Forest (TSF) algorithm has been published, which outperformed 1-NN DTW with a “computational complexity linear in the time series length” [88]. However, on average, this algorithm is quite weak in terms of accuracy [88, 89]. Two other interval-based algorithms, the *Time Series Bag of Features* and a classifier using similarities based on local autopatterns [90] “have been shown to be no more accurate than TSF on average while being considerably slower” [89]. Alternative approaches include Random Interval Spectral Ensemble (RISE), “a tree ensemble that extracts spectral features from intervals” [89] and its successor Contract Random Interval Spectral Ensemble (c-RISE). RISE builds each tree on a distinct set of features extracted with a Fourier transform, autocorrelation and partial autocorrelation. It is part of the meta ensemble Hierarchical Vote Collective of Transformation-based Ensembles (HIVE-COTE) (see Section 2.4.6) and has “a run time complexity of $\mathcal{O}(n \cdot m^2)$, where m is the time series length and n the number of train cases” [91]. The algorithm c-RISE “adaptively estimates the time taken to build each tree in the ensemble”, which makes it “more effective than the static approach of estimating the complexity before executing” [91].

2.4.6. Ensemble methods

Ensemble methods consist of multiple ML algorithms to obtain a better performance than the expected performance of any of the constituent ML algorithms alone. One can distinguish between different types of ensemble methods as described in the following.

2.4.6.1. Distance-based ensembles

In 2015, an ensemble algorithm that significantly outperformed individual distance-based classifiers, the proportional *Elastic Ensemble (EE)*, has been published. The authors considered it “the first ever classifier to significantly outperform [1-NN DTW] (see Section 2.4.2)” on TSC tasks [92]. More recently, *Proximity Forest (PF)*, an ensemble of trees scaling "quasi-linearly with the quantity of training data", has been introduced. Even though it has been shown to be "significantly more accurate than EE", it did not achieve superior accuracies in comparison to *Collective of Transformation-based Ensembles (COTE)* [93].

2.4.6.2. Transformation-based ensembles

COTE has been introduced in the same year as EE and demonstrated superior accuracy [94]. Flat Collective of Transformation-based Ensembles (Flat-COTE), a variant of the COTE model, "combining predictions of 35 individual classifiers built on four representations of the data into a flat hierarchy", has been shown to be a very "effective ensembling strategy". It yields better results in comparison to other COTE variants and certain deep learning strategies (see Section 2.4.7) [95].

2.4.6.3. Hybrid ensembles

HIVE-COTE, an ensemble including EE, a *Shapelet Transform (ST)* ensemble (see Section 2.4.4), a Time Series Forest ensemble, BOSS (see Section 2.4.3) and a newly developed spectral ensemble has been shown to be "significantly more accurate than Flat-COTE [95] and [represented] a new state-of-the-art for TSC" on data from the UCR archive in 2018 [96]. However, the high accuracy of HIVE-COTE comes at the cost of a comparatively high computational training complexity, which makes it "slow, even for smaller [datasets and] intractable for large datasets" [83]. The more scalable ensemble method *Time Series Combination of Heterogeneous and Integrated Embedding Forest (TS-CHIEF)* [97] builds on PF and incorporates dictionary-based and interval-based splitting criteria. It has a "quasilinear training complexity in the number of training examples but quadratic training complexity in time series length". It has also been shown to be slightly less accurate than much faster methods, such as *Random Convolutional Kernel Transform (ROCKET)* (see Section 2.4.7) [83]. A recent improvement of HIVE-COTE replaced its TSF component with *Canonical Interval Forest (CIF)*, which embeds *set of 22 CAnonical Time-series CHaracteristics (catch22)* [98] (see Section 2.5.1) in an adaptation of TSF. This new classifier termed HIVE-COTE with CIF (HC-CIF) has been shown to be "significantly more accurate" than HIVE-COTE [89]. Another HIVE-COTE derivative is *HIVE-COTE with Temporal Dictionary Ensemble (HC-TDE)*, which replaces the BOSS model with a dictionary-based classifier using a new approach for "constructing ensemble members based on an adaptive Gaussian process model". It has been shown to achieve state-of-the-art accuracies when it was published in 2020 [99].

2.4.6.4. Deep learning ensembles

1-D deep learning models for TSC have only recently become the subject of research, with the first architectures developed over the last few years. Section 2.4.7 discusses those models in detail. An ensemble of deep CNNs is *InceptionTime*, which has been published in 2019. This architecture consists of five deep learning models and slightly outperforms HIVE-COTE, while being two orders of magnitude faster at the same time [100]. Even though the speed and accuracy of InceptionTime was very competitive at the time it was published, it has since been shown to yield slightly less accurate results in comparison to more recent 1-D CNNs, such as ROCKET (see Section 2.4.7), while being considerably slower [83].

2.4.6.5. Gradient boosting machines

GBMs are ML models predicting outcomes based on the output of an ensemble of weak models, typically Gradient Boosted Decision Trees (GBDTs). Popular GBDTs representatives are *Extreme Gradient Boosting (XGBoost)* [101], *Light Gradient Boosting Machine (LightGBM)* [102] or *CatBoost* [103] (see Section 2.7.4.4 for a more detailed discussion). In a 2018 publication, a XGBoost classifier based on features extracted from ECG signals has been shown to outperform RFs in detecting cardiac anomalies [104], while a paper published in 2019 demonstrated that a classifier, which was based on a LightGBM, has been able to classify EEG signals better than traditional classifiers, such as SVMs or CNNs [105]. More recently, a comprehensive review from 2020 investigated the CatBoost algorithm and found it to be a “good candidate for ML implementations involving [big data]” and recommended researchers to use it with datasets “that are heterogeneous and have categorical features”. This review also found a high sensitivity to hyper-parameter settings of this method [106].

2.4.7. (Deep) artificial neural networks for 1-D signals

An alternative approach for TSC are 1-D ANNs (see Section 2.7.4.6 for a detailed explanation). Figure 2.11 shows a taxonomy of several Artificial Neural Network (ANN) methods for TSC [61].

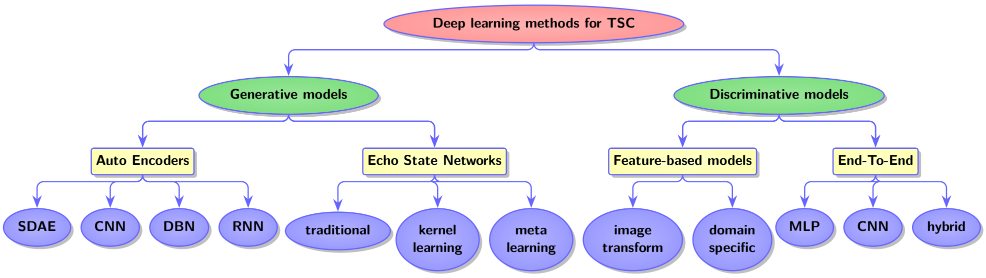


Figure 2.11.: Taxonomy of ANN methods for TSC (adapted from [61]).

A strategy, first published in 1997, to recognize patterns was Long

Short-Term Memory (LSTM) with which it became possible to “solve many previously unlearnable DL tasks”. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points, but also entire sequences of data. Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs) “won several international pattern recognition competitions and set numerous benchmark records on large and complex datasets” [107]. A paper published in 2003 discussed the possibility of applying Radial Basis Function Neural Networks (RBFNNs) for the classification of bioacoustic time series [108]. RBFNNs classify an input feature vector $x^{\mu} \in \mathbb{R}^{d}$ into one of l different classes consisting of d input neurons, l output neurons and a layer of k nonlinear RBF neurons. A publication from 2018 challenged LSTM RNNs as a common default model for sequence modeling tasks by arguing “that the common association between sequence modeling and (LSTM) recurrent networks should be reconsidered and [CNNs] should be regarded as a natural starting point for sequence modeling tasks” instead [109]. In 2016, “a simple but strong baseline for [TSC] from scratch with deep neural networks” had already been proposed by introducing end-to-end models that do not require any heavy preprocessing or feature crafting. This baseline included the 1-D models Multilayer Perceptron (MLP), Fully Convolutional Network (FCN) and Residual Network (ResNet), which were already able to yield “comparable or better” results to state-of-the-art models at that time [110]. A comprehensive review, published in 2019, investigated several 1-D CNNs and concluded that deep neural networks are able to significantly outperform 1-NN DTW and to achieve results not significantly different than results from COTE or HIVE-COTE (see Section 2.4.6) [61]. ROCKET, which was published in the same year, exploits simple linear classifiers and random convolutional kernels. It has been shown to be much faster than other algorithms, such as PF, InceptionTime, TS-CHIEF or HIVE-COTE, yielding comparable and, in many cases, superior accuracies [83]. A publication from 2020 introduced *Omni-Scale 1D-CNN*, which has been evaluated on the UCR archive yielding state-of-the-art performance on some indicators [111]. More recently, a derivative of ROCKET, termed Minimally Random Convolutional Kernel Transform (MiniROCKET), has been introduced. This method has been shown to be “up to 75 times faster on larger datasets”, while achieving roughly the same accuracies as ROCKET [112]. Even more recently, Minimally Random Convolutional Kernel Transform with Multiple Features (MultiRocket) has been introduced, which significantly improves the accuracy of MiniROCKET and ROCKET with some additional computational expenses. MultiRocket has been shown to be “the current most accurate univariate TSC algorithm on the [datasets] in the UCR archive. While approximately 10 times slower than [MiniROCKET], MultiRocket is still much faster than other state-of-the-art [TSC] algorithms”, such as HIVE-COTE, HC-CIF or HC-TDE [113].

2.4.8. (Deep) artificial neural networks for images

Section 2.4.7 introduces 1-D ANNs, but time series can also be transformed into data with higher dimensionality, such as 2-D images, before classification.

A common 2-D representation of 1-D signals are Recurrence Plots (RPs), which were introduced in 1987 [114] and provide “a way to visualize the periodic nature of a trajectory through a phase space and [enable the investigation] of certain aspects of the m-dimensional phase space trajectory” [115]. In a 2018 publication, RPs have been combined with CNNs to perform TSC tasks with higher accuracies than selected methods on datasets from the UCR archive. However, not all available datasets have been included in this study [115]. In the same year, a CNN using RPs of 1-D tri-axial acceleration signals has been applied to distinguish between different human activities [116]. A year later *Timage*, a TSC pipeline making use of transfer learning in ANNs and RPs, has been introduced. Timage was evaluated on the 2018 release of the UCR archive but failed to improve the state-of-the-art in terms of accuracy or speed [62]. Another method introduced in 2020 makes use of a FCN in combination with Multi-scale Signed Recurrence Plots (MS-RPs) to achieve superior performances in comparison to other state-of-the-art classifiers on signals from preselected datasets from the UCR archive. MS-RPs build upon RPs and enrich them by adding phase space dimensions and time delay embeddings [117].

A publication from 2015 encodes times series as Gramian Angular Fields (GAF), Markov Transition Fields (MTF) and a combination of both approaches to use these newly created images as input data for CNNs. A GAF represents time series in a polar coordinate system instead of the typical Cartesian coordinates. In the resulting Gramian matrix, each element is the cosine of the summation of angles. A MTF builds the Markov matrix of quantile bins after discretization and encodes the dynamic transition probability in a quasi-Gramian matrix. Combined GAF and MTF images used as input data for a CNN were competitive to state-of-the art approaches, such as 1-NN DTW, ST, BOP, SAXVSM, at that time [118].

Another approach for TSC is transforming the input signals to *spectrograms* first, which provide a visual representation of the frequency spectrum of a signal as it varies in time [119].

2.5. Features

Different data types can be used as input for different kinds of ML models. Instead of raw data, processed data, such as feature vectors are also a possibility. Using feature vectors instead of raw data can, in some cases, prove to be advantageous. Before the advent of end-to-end learning models, the performance of all ML methods was heavily dependent on the choice of data representation (i.e. features) on which they were applied [120]. Today, features are still an important part of ML pipelines. They are often computed as a preprocessing step or as part of the model itself. The following sections discuss feature engineering (see Section 2.5.1), feature extraction (see Section 2.5.2, feature learning (see Section 2.5.3), feature selection (see Section 2.5.4) and feature scaling (see Section 2.5.5) approaches in detail. This work also exploits extracted features instead of raw input signals, which can lead to improved accuracies or faster run times in certain cases. Domain specific knowledge (DSK) of the behavior of US signals under certain circumstances (i.e. expected amount of reflection or transmission based on the examined tissue) can be a crucial part of good ML model performances.

2.5.1. Feature Engineering

Before the advent of end-to-end ML models, a lot of effort went into the design of preprocessing pipelines and data transformations creating data representations suitable to achieve high accuracies. This important but labor intensive design process is called *feature engineering* [120] and is of crucial importance for algorithms unable to extract discriminative information from the data on their own. Feature engineering often incorporates DSK but ML can also leverage general priors that are not task specific.

In contrast to general-purpose priors, DSK can also be an important cornerstone of ML pipelines. For example, US specific properties, such as attenuation factors, speed of shear waves or time of flight values can be used in feature vectors in addition to statistical features (see Section 2.5.2). Many of those features require knowledge about the used sample frequency or amplitude peaks expected in the signals whose presence, in turn, often depends on the behavior of the US signals in certain types of tissues.

2.5.2. Feature Extraction

Once promising features have been engineered (see Section 2.5.1), they can be extracted from the input data. This can either be done by writing custom functions or by relying on external software libraries. Libraries such as *highly comparative time-series analysis (hctsa)* [121], *catch22* [98] or *Time Series Feature Extraction Library (TSFEL)* [66] are very useful for this purpose and are also used in this work. *hctsa* can extract more than 7,000 features from time series using Matlab, while *catch22* is a collection of 22 time-series features that can be computed with Python, R, Matlab or Julia. The *catch22* features are a subset of the *hctsa* features. TSFEL is a Python library, which can compute 60 different features from temporal, statistical and spectral domains. Table 2.4 provides an overview of all features provided by TSFEL. See Section 2.3.3 for a detailed description how those features are computed.

2.5.3. Feature Learning

Feature learning is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification

Statistical domain	Temporal domain	Spectral domain
ECDF	Absolute energy	FFT mean coefficient
ECDF Percentile	Area under the curve	Fundamental frequency
ECDF Percentile Count	Autocorrelation	Human range energy
Histogram	Centroid	LPCC
Interquartile range	Entropy	MFCC
Kurtosis	Mean absolute diff	Max power spectrum
Max	Mean diff	Maximum frequency
Mean	Median absolute diff	Median frequency
Mean absolute deviation	Median diff	Power bandwidth
Median	Negative turning points	Spectral centroid
Median absolute deviation	Peak to peak distance	Spectral decrease
Min	Positive turning points	Spectral distance
Root mean square	Signal distance	Spectral entropy
Skewness	Slope	Spectral kurtosis
Standard deviation	Sum absolute diff	Spectral positive turning points
Variance	Total energy	Spectral roll-off
	Zero crossing rate	Spectral roll-on
	Neighbourhood peaks	Spectral skewness
		Spectral slope
		Spectral spread
		Spectral variation
		Wavelet absolute mean
		Wavelet energy
		Wavelet standard deviation
		Wavelet entropy
		Wavelet variance

Table 2.4.: Overview of all TSFEL features.

from raw data. This replaces manual feature engineering (see Section 2.5.1) and allows an algorithm to learn the features on its own and use them to perform a specific task. Feature learning can be either *supervised* or *unsupervised*, i.e. with or without labeled input data. Approaches for supervised feature learning include dictionary learning (see Section 2.4.3) or ANNs (see Section 2.4.7). Examples for unsupervised feature learning strategies, whose goal is to discover low-dimensional features that capture some structure underlying the high-dimensional input data, include k-means clustering, Principal Component Analysis (PCA) (see Section 2.7.3.1) and other Dimensionality Reduction Techniques (DRTs) (see Section 2.7.3).

2.5.4. Feature Selection

Feature selection is the process of selecting a subset of suitable features for ML models. Feature selection techniques are used for the following reasons [122]:

- facilitating data visualization and data understanding
- reducing measurement and storage requirements
- reducing training and utilization times

- defying the curse of dimensionality to improve prediction performance (see Section 2.7.1.1)

Commonly, feature ranking algorithms are used to select features [122].

2.5.5. Feature Scaling

Feature scaling or *data normalization* is a pre-processing step in which the input data is either scaled or transformed to make sure that all features contribute equally in terms of values. As a good performance of ML algorithms depends upon the quality of the input data, feature scaling is very crucial [123]. Various approaches exist, that can be categorized into the following strategies [123]:

- Mean and standard deviation based methods
- Minimum-Maximum value based methods
- Decimal scaling normalization
- Median and median absolute deviation normalization
- Tanh based normalization
- Sigmoidal normalization

The popular Python library *scikit-learn* provides many methods to scale the input data [124].

2.6. Digital Signal Processing

DSP is the use of digital processing to perform a wide variety of signal processing operations. The digital signals processed in this manner are a sequence of numbers that represent samples of a continuous variable in domains such as time, space, or frequency.

DSP applications include a wide range of fields. However, this work focuses solely on processing biomedical signals, which has been considered to be “one of the most important [visualization and interpretation] methods in biology and medicine” [125]. DSP methods can either be performed on raw input signals or on pre-processed features. They are often implemented in microcontrollers or dedicated embedded DSP chips.

A large variety of different DSP methods for many different applications exists. This section provides a brief overview of different domains DSP can be performed in. Digital 1-D signals can be analyzed in the *time domain* (i.e. how the signal behaves with respect to time), while digital 2-D images can be directly analyzed in the *space domain* (i.e. pixel locations). By applying the Fourier transform (see Section 2.3.2.1), the time or space information can be transformed to the *frequency domain* by extracting magnitude and phase components of each frequency. Frequency domain analysis is helpful to examine signal properties or to filter the signals.

The analysis and processing of very large datasets (i.e. *big data*), the extraction of promising features and the determination of appropriate thresholds pose significant challenges for classical DSP approaches due to high costs with respect to memory and computing resources. More sophisticated approaches extending classical signal processing techniques, like discrete signal processing on graphs, exist to circumvent these limitations [126] but further research is still needed to combine graph signal processing with existing techniques in a sensible way [127]. Section 2.7 provides a detailed description of alternative ML approaches to tackle the limitations of classical DSP methods.

2.7. Machine Learning

ML algorithms are algorithms that improve their performances automatically through experience and by the use of data. Figure 2.12 shows a mindmap categorizing ML as a sub-field of Artificial Intelligence (AI). It is obvious that ML is not equal to AI and only constitutes a single building block. ML techniques build a model based on sample data, known as "training data", to create predictions without being given any explicit instructions and are used in a wide variety of applications where it is difficult or infeasible to use more traditional algorithms, such as approaches based on DSP .

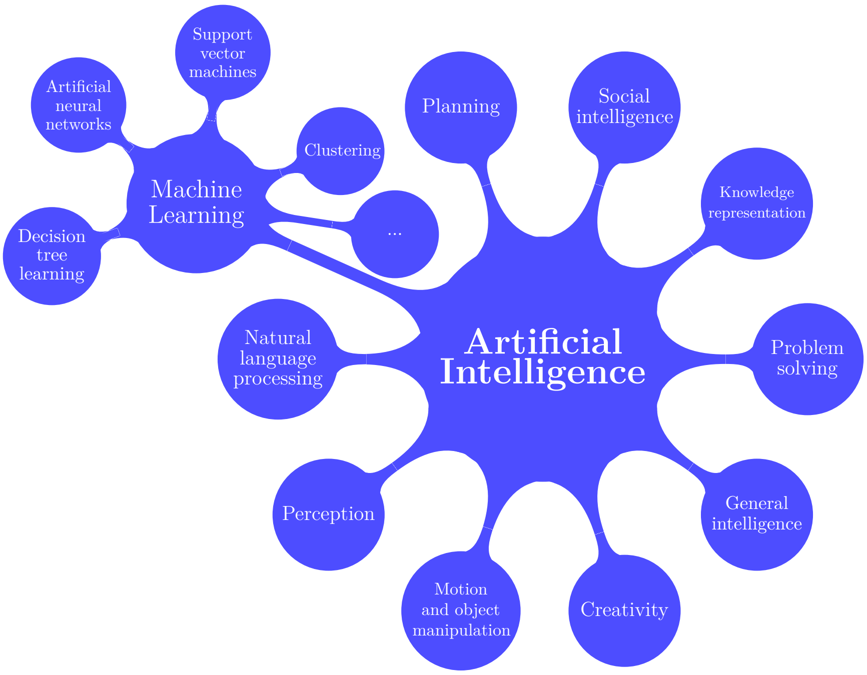


Figure 2.12.: Mindmap showing the relationship between artificial intelligence and machine learning.

This section restricts itself to providing details concerning ML for TSC as a complete comprehensive approach would be way out of scope for this work. Section 2.7.1 introduces the topic, while Section 2.7.3 describes DRTs used for preprocessing or visualization purposes. Section 2.7.4 introduces a variety of different models used for TSC and Section 2.7.5 distinguishes different approaches for model evaluation.

2.7.1. Introduction

A ML model is learning if it improves its performance on future tasks after making observations about the world. Learning strategies can be categorized into *unsupervised learning*, *reinforcement learning*, *supervised learning* and *semi-supervised learning* [128]. In unsupervised learning, the ML model learns patterns without any explicit feedback or labeled input data. The most common unsupervised learning task is *clustering*, which aims to categorize the inputs into clusters without having any knowledge about them beforehand. In reinforcement learning the ML model learns from a series of rewards or punishments. Based on the feedback, the model learns to discriminate between steps leading to rewards and steps leading to punishments. Supervised learning strategies rely on labeled input data. The annotation is usually done by human experts. In semi-supervised learning, the model is presented with a few labeled examples and has to find ways to expend its knowledge to unlabeled examples [128]. In this work, all input data has been labeled and is to be categorized into two different classes (e.g. *relaxed muscle / fatigue muscle* as described in Chapter 3). Thus, only supervised models for binary classification are examined further.

2.7.1.1. Curse of Dimensionality

An important issue for most ML tasks is the *curse of dimensionality*. This concept refers to the data sparsity occurring when moving to higher dimensions. The volume of the space represented grows so quickly that the data cannot keep up and thus becomes sparse. This phenomenon can be addressed using polynomial curve fitting. If there are D input variables, the general polynomial with coefficients up to order 3 would take the following form [77]:

$$y(x,w) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k. \tag{2.14}$$

As D increases, the number of independent coefficients grows proportionally to D^3. One way to address the increasing complexity of a ML model for high dimensional data is to deploy DRTs (see Section 2.7.3) to reduce the dimensionality. These methods can either be deployed as preprocessing steps before using the data in any model or as a way to intuitively visualize high dimensional data with only two or three dimensions.

2.7.1.2. Bias-variance trade-off

The *bias-variance trade-off* is a fundamental principle for understanding the generalization of ML models. *Bias* describes the difference between the average prediction of the ML model and the groundtruth. *Variance* is a measure for the

sensitivity to fluctuations in the training set [129]. It measures the variability of the ML model prediction for a value which provides information about the spread of the input data.
Until recently, it was common knowledge that model variance increases and bias decreases monotonically with model complexity but recent work calls this belief into question for ANNs and other over-parameterized models. For those ML models it is often observed that larger models generalize better [129]. A high bias can cause an algorithm to miss relevant relations between features and target outputs, while a high variance may result in an algorithm being affected by random noise. The former is known as *underfitting*, while the latter is called *overfitting*.

2.7.2. Software

A large variety of different programming languages, frameworks and libraries exist to perform ML in general and TSC in particular. This section provides a brief overview of available technology and software in this field. This work makes use of all software solutions described in the following. Section 2.7.2.1 provides an overview of general Python ML frameworks and libraries, while Section 2.7.2.2 covers TSC libraries in particular. Section 2.7.2.3 introduces software used for feature processing.

2.7.2.1. Machine Learning

The programming language *Python* is a popular choice for data science and ML tasks. The free and open-source ML libraries *TensorFlow* and *PyTorch* [130, 131] have a particular focus on training and inference of DL networks and form the foundation of many DL models throughout this work. Another library used for this work is *Keras*, a high level interface for TensorFlow. It provides a large collection of commonly used neural network building blocks such as layers, objectives, activation functions, optimizers and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code [132]. Another crucial open source library finding usage in this work is *scikit-learn*, which builds on *NumPy* [68], *SciPy* [69] and *matplotlib* [133]. scikit-learn provides various tools for supervised learning, unsupervised learning, model fitting, data preprocessing, model selection, model evaluation, data visualization and many other utilities [124].

2.7.2.2. Time series classification

For TSC in particular, several sophisticated Python libraries exist. One example is *sktime*, which provides a unified interface for multiple time series learning tasks. This includes time series classification, regression, clustering, annotation and forecasting. It comes with time series algorithms and scikit-learn [124] compatible tools to build, tune and validate time series models [134]. sktime provides efficient implementations of many TSC algorithms (see Section 2.4). For classification tasks using the 1-NN DTW algorithm (see Section 2.7.4.2), this work relies on the efficient *DTAIDistance* library, which offers a pure Python implementation and a faster implementation in *C* [135].

2.7.2.3. Features

Several Python libraries provide the possibility to automatically extract, learn or create features (see Section 2.5) from time series inputs. One possibility is the Python package *tsfresh*, which also contains methods to evaluate the respective explaining power and importance of certain features [70]. The *TSFEL* Python library can extract over 60 different features from statistical, temporal and spectral domains [66]. The *catch22* Python library is able to compute time series features, such as "linear and non-linear autocorrelation, successive differences, value distributions and outliers, and fluctuation scaling properties" [98].

2.7.3. Dimensionality Reduction Techniques

DRTs are techniques to transform high-dimensional input data into a low-dimensional representation, such that the latter retains meaningful properties of the original data. These properties are ideally close to the intrinsic dimension of the original data. Due to the curse of dimensionality (see Section 2.7.1.1), the raw data are often sparse and analyzing the data is usually computationally expensive. Dimensionality reduction is common in fields that deal with large numbers of observations and/or large numbers of variables.

Traditional DRTs, such as PCA or Linear discriminant analysis (LDA) "are linear techniques that focus on keeping low-dimensional representations of dissimilar data points far apart". For non-linear manifolds it is usually "more important to keep the low-dimensional representations of very similar high-dimensional data points close together, which is typically not possible with a linear mapping" [136]. DRTs are especially important in the context of mobile approaches. Dimensionality reduced signals can significantly speed up the training and inference computations of ML models in general and ML models for mobile devices in particular. This also impacts the energy consumption of those devices. Even though many different methods exist, this section restricts itself to two methods used in this work. Section 2.7.3.1 introduces the linear technique PCA, while Section 2.7.3.2 discusses the non-linear method t-distributed stochastic neighbor embedding (t-SNE).

2.7.3.1. Principal component analysis

The English mathematician and biostatistician Karl Pearson (1857-1936) originally formulated *PCA* in 1901 as a minimization of the sum of squared residual errors between projected data points and the original data. PCA is defined as an orthogonal linear transformation that transforms the data to a

new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate and so on [137]. An extension to PCA called *kernel PCA* exists, which allows to apply the concept as nonlinar transformation.
Sections 3.4.1.1, 3.4.3.1, 4.4.1.1 and 5.4.1.1 show how the PCA technique can be used in this work to gain a better understanding of the high-dimensional data distribution of the underlying input data.

2.7.3.2. t-distributed stochastic neighbor embedding

An alternative Dimensionality Reduction Technique (DRT) is t-SNE, which was presented by the Dutch research scientist Laurens van der Maaten and the British-Canadian cognitive psychologist and computer scientist Geoffrey Hinton in 2008 [136]. This technique is based on *Stochastic Neighbor Embedding* and "is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales". t-SNE "minimizes the *Kullback-Leibler* divergence between the joint probabilities p_{ij} in the high-dimensional space and the joint probabilities q_{ij} in the low-dimensional space" [136]. The Kullback-Leibler divergence is a measure of how one probability distribution is different from a second probability distribution.
Sections 3.4.1.2, 3.4.3.2, 4.4.1.2 and 5.4.1.2 show how the t-SNE technique can be used in this work to gain a better understanding of the high-dimensional data distribution of the underlying input data.

2.7.4. Classification Models

Section 2.4 provides an overview and a taxonomy of different classification approaches for TSC. This section aims to provide greater detail for the ML classification models used in this work. Section 2.7.4.1 introduces *LR*, Section 2.7.4.2 discusses *1-NN DTW* and Section 2.7.4.3 provides an overview of *DTs*. Section 2.7.4.4 provides an overview of *GBMs*, while Section 2.7.4.5 discusses *SVMs*. Section 2.7.4.6 provides details of *ANNs*. This work deliberately does not make use of certain techniques or models. Section 2.7.4.7 explains the reasoning behind omitting certain models in greater details.

2.7.4.1. Logistic Regression

The logistic function has convenient mathematical properties and is defined as follows [128]:

$$f(z) = \frac{1}{1 + e^{-z}}. \tag{2.15}$$

If $h_w(x)$ is a hypothesis function that is not differentiable and is a discontinuous function of its inputs and its weights w, the process of fitting the weights of the model

$$h_w(x) = \frac{1}{1 + e^{(-w \cdot x)}} \tag{2.16}$$

to minimize loss on a dataset is called *LR*. LR is comparatively slow to converge for linearly separable data but behaves much more predictably than other methods. If the data are noisy and non-separable, LR converges “far more quickly and reliably” and has become “one of the most popular classification techniques” for a variety of different applications [128].

2.7.4.2. 1-Nearest Neighbor with Dynamic Time Warping

The theoretical foundations for the *1-NN DTW* algorithm were already laid in a publication from 1959 [138]. Subsequently, the original algorithm and more sophisticated extensions have been used for decades in the fields of pattern recognition, speech recognition, Natural Language Processing (NLP) and TSC [63, 139]. 1-NN DTW was, until recently, a very popular benchmark for TSC and considered to be “hard to beat” [81]. 1-NN DTW is a distance-based classifier (see Section 2.4.2), that substitutes the *Euclidean distance* with an alternative distance metric called *dynamic time warping* [63]. The naïve 1-NN DTW implementation has a time complexity of $\mathcal{O}(N \cdot M)$ for the comparison of two time series with length N and M [139].

2.7.4.3. Decision Trees

Decision tree learning is a predictive modeling approach using a decision tree to go from observations about an item (represented in the branches) to conclusions about the item’s target value (represented in the leaves). The leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Tree models can either be classification trees if the target variables are a discrete set of values or regression trees if the target variables are continuous values, such as real numbers. Early roots of this algorithm can be traced back to the work of Ronald Fisher, who used a classification tree model to introduce LDA in 1936 [140]. The first regression tree algorithm in the modern sense, called *Automatic Interaction Detection* was published in 1963. Initially this work did not attract much interest in the research community but an approved algorithm called *Classification And Regression Trees* revived research interest in 1984 [140].

A binary classification tree “reaches its decision by performing a sequence of tests. Each internal node in the tree corresponds to a test of the value of one of the input attributes A_i, and the branches from the node are labeled with the possible values of the attribute, $A_i = v_{ik}$. Each leaf node in the tree specifies a value to be returned by the function” [128]. Figure 2.13 shows a complex sample decision to predict advanced liver fibrosis in chronic hepatitis C patients (see Section 5.1). In the figure, liver markers include alpha-fetoprotein (AFP), aspartate aminotransferase (AST) and small fragments of the cytoplasm called platelets. S represents the fibrosis score of the patient. If the final value of $S \geq 0$, then the patient has an advanced fibrosis and vice versa [141].

Decision trees are rarely used on their own for TSC. However, they provide the underlying foundation for many more advanced approaches, such as RFs

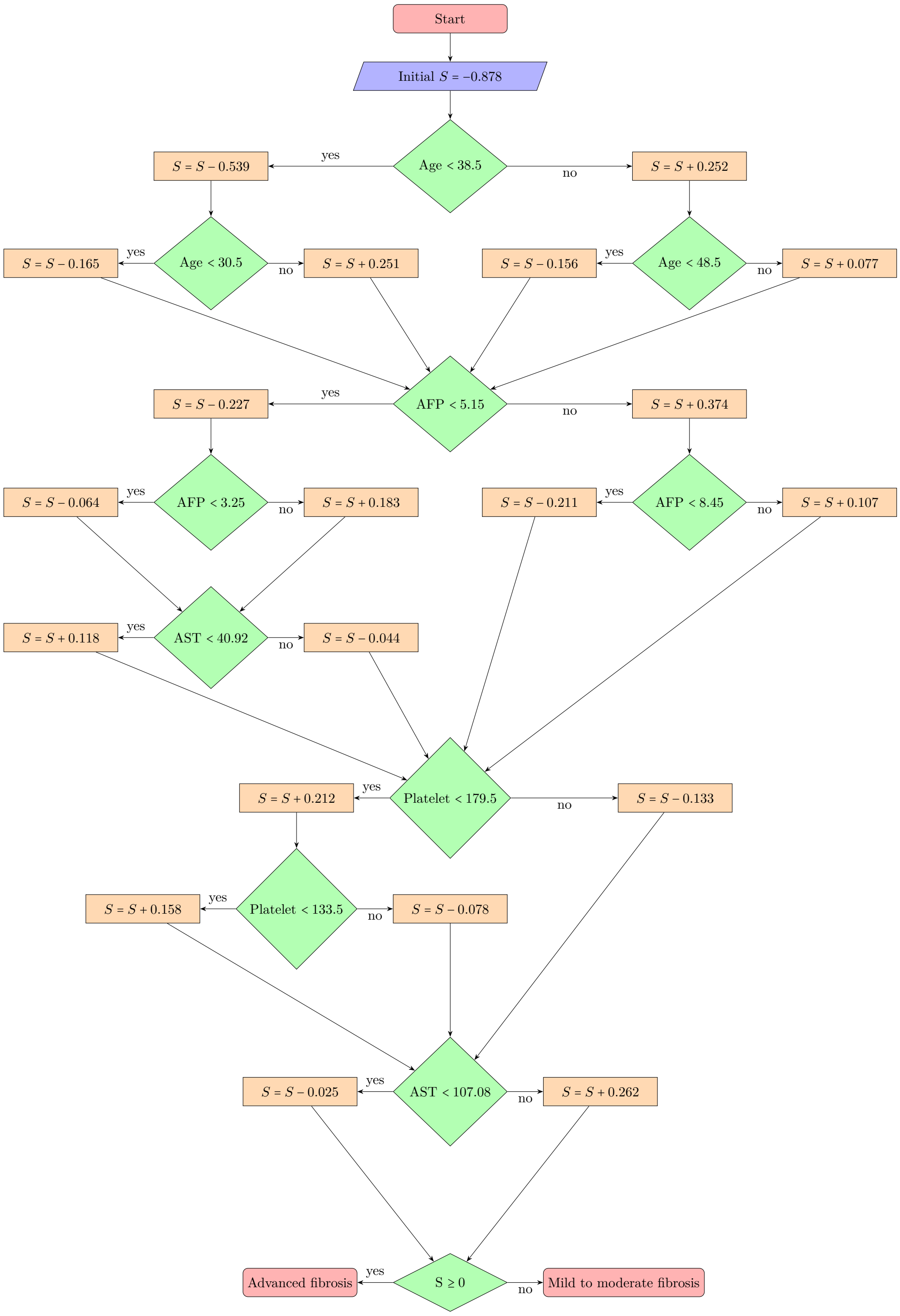


Figure 2.13.: Sample decision tree to predict advanced liver fibrosis in patients with chronic hepatitis C. If the final value of $S \geq 0$, the patient has an advanced fibrosis and vice versa (adapted from [141]).

or GBMs (see Section 2.7.4.4).

2.7.4.4. Gradient Boosting Machines

GBMs are ML models that make predictions based on an ensemble of weak prediction models, typically DTs (see Section 2.7.4.3). When a decision tree is used as a weak learning algorithm, the resulting ensemble is called GBDTs. Section 2.4.6.5 provides an overview of current and popular GBDT methods. GBDTs train an ensemble model of decision trees in sequence. In each iteration, GBDTs learn decision trees by adjusting negative gradients [102]. The main computational cost of GBDTs lies in the creation of the decision trees, while the most time-consuming part in learning a decision tree is to find the most suitable splitting locations. The idea of gradient boosting occurred based on the observation that boosting can be interpreted as an optimization approach with a suitable cost function [142].

Presumably, the most popular GBDT choices are *XGBoost* [101], *LightGBM* [102] and *CatBoost* [103]. This section describes the main differences between these methods.

Before learning any tree structure, all GBMs first create feature-split pairs for all features, that are based on histograms and used as possible node splits. *XGBoost* does not make use of any weighted sampling techniques to optimize the speed of the splitting process. LightGBM enables *Gradient-based One-Side Sampling (GOSS)*, which keeps instances with large gradients (e.g. gradients that are larger than a given threshold or among the top percentiles) and only randomly drops instances with small gradients [102]. By default, CatBoost makes use of the *Minimal Variance Sampling (MVS)* algorithm to evaluate candidate splits, when each next decision tree is constructed. The weighted sampling then happens at tree-level and not at the split-level. The observations for each boosting tree are sampled in a way that maximizes accuracies. Recently, efforts have been described to substitute the GOSS method in LightGBM by MVS to achieve faster and more accurate results [143].

The initial implementation of *XGBoost* deployed a level-wise tree growth [101], while LightGBM performs leaf-wise tree growth [102]. The former populates the tree with a focus on width, while the latter populates the tree with a focus on depth. By default, CatBoost grows a balanced tree [103].

While LightGBM and XGBoost allocate missing values to the tree branch that reduces the loss in each split [101, 102], CatBoost either processes missing values as minimum or maximum value for a feature [103].

As there are no publications providing clear evidence for the general superiority of any of the GBM methods mentioned above, all are deployed in this work to allow for a fair comparison.

2.7.4.5. Support Vector Machines

SVMs are ML models whose theoretical roots can be traced back to work done in 1962, which was first published in 1964 [144]. “The extraordinary generalization capability of SVM, along with its optimal solution and its discriminative power, has attracted the attention of data mining, pattern recognition and [ML] communities in the last years. SVM has been used as a powerful tool for solving practical binary classification problems [and] it has been shown that SVMs are superior to other supervised learning methods. Due to its good theoretical foundations and good generalization capacity, in recent years, SVMs have become one of the most used classification methods" [145]. The original SVM classifier, used for classification and regression tasks, separates datasets with two classes 1 and 2 by finding an optimal hyperplane. This linear hyperplane is defined as [60]:

$$W^T x + b = \begin{cases} \geq 1, \text{class } 1 \\ \leq -1, \text{class } 2, \end{cases} \tag{2.17}$$

where $x \in R^{n \times 1}$ is the input vector with n features, $W \in R^{n \times 1}$ is a weight vector and b is a bias term. Figure 2.14 shows a sample linear hyperplane of a binary linear two-dimensional SVM used to separate data points from two distinct categories. The black data points belong to class 1, while the empty black circles belong to class 2. The red data points and the empty red circles are located on the border of the corresponding classes. According to equation 2.17 those belong to class 1 and 2 respectively but the definition could be changed to associate those data points to a different class. The separation margin between the classes is given by $\frac{2}{||W||}$, where $||W||$ is the 2-norm of the weight vector, which is being minimized by the SVM algorithm to maximize the separation margin. The maximum margin can be found by solving the following quadratic optimization problem [60]:

$$\min \frac{1}{2} ||W||^2, \tag{2.18}$$

subject to $y_i(W^T x_i + b) \geq 1$ with x_i being the ith input vector and $y_i \in \{1, 2\}$ being the corresponding label for x_i. If the input data are linearly separable,

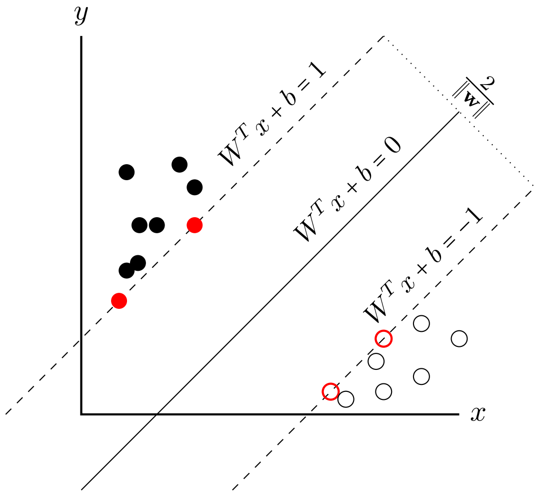


Figure 2.14.: A two-dimensional feature space with optimal separating hyperplane.

a simple SVM as depicted in Figure 2.14 is a suitable choice. To classify non-linearly separable data, "the input vector $x \in \mathbb{R}^n$ [can be transformed] into vectors $\Phi(x)$ of a highly dimensional feature space F". This transformation is defined as follows [145]:

$$x \in \mathbb{R}^n \to \Phi(x) = [\phi_1(x), \phi_2(x), \cdots, \phi_n(x)]^T \in \mathbb{R}^f. \tag{2.19}$$

The linear classification can then be solved in this feature space. "The decision rule can be evaluated using dot products" as follows [145]:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i \langle \phi(x_i) \cdot \phi(x) \rangle + b. \tag{2.20}$$

"If there is a way to calculate the product $\langle \phi(x_i) \cdot \phi(x) \rangle$ in the feature space directly as a function of the original input data, this makes it possible to join the two necessary steps to build a non-linear learning machine" called a kernel function. "Some of the most used kernel functions are" [145]:

1. **Linear kernel:** $K(x_i, x_j) = (x_i \cdot x_j)$
2. **Polynomial kernel:** $K(x_i, x_j) = (x_i \cdot x_j + 1)^p$
3. **Gaussian kernel:** $K(x_i, x_j) = e^{\frac{-\|x_i - x_j\|^e 2}{2\sigma^2}}$
4. **RBF kernel:** $K(x_i, x_j) = e^{-\gamma(x_i - x_j)^2}$
5. **Sigmoid kernel:** $K(x_i, x_j) = tanh(\eta x_i \cdot x_j + v)$

Whether a certain kernel is better or worse depends on the specific application but previous works have generally concluded that "the polynomial and the Gaussian RBF function are the best option for acoustic signals" [145].

2.7.4.6. Artificial neural networks

ANNs are algorithms based on nodes that loosely model neurons in a biological brain. These nodes or units are connected by directed links. A link from unit i to unit j serves to propagate the activation a_i from i to j. Each link also has a numeric weight $w_{i,j}$ associated with it, which determines the strength and sign of the connection. Each unit has a dummy input $a_0 = 1$ with an associated weight $w_{0,j}$. "Each unit j first computes a weighted sum of its inputs" [128]:

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i. \tag{2.21}$$

Then it applies an *activation function* g to this sum to derive the output [128]:

$$a_j = g(in_j) = g(\sum_{i=0}^{n} w_{i,j} a_i). \tag{2.22}$$

Activation functions

Activation functions are the primary decision-making units of ANNs. They evaluate the node outputs and are crucial for the general performance. The choice of a suitable activation function is critical as it significantly influences the performance of the whole neural network. In practice, only a few functions are commonly used and well studied for ANN analysis. Figure 2.15 illustrates some examples of popular choices, such as the *sigmoid function (a)*, *tanh (b)*, the *ReLU function*, the *leaky ReLU function* and the *swish function* [146]. The sigmoid function is defined as follows [146]:

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{2.23}$$

The sigmoid function is nonlinear and has a clear tendency of bringing the Y values to either end of the curve, which results in clear prediction distinctions. Even though it is bound by its range $(0,1)$, this function is not without disadvantages. It can be prone to the vanishing gradient problem, which is encountered if the gradient becomes so small that its value effectively prevents the weights from changing.
The tanh function is defined as follows [146]:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{2.24}$$

This function has the range $(-1,1)$, which allows negative outputs as well. The derivatives of tanh are significantly larger than the derivatives of the sigmoid function. However, tanh might still be impacted by the vanishing gradient problem [146].
The basic Rectified Linear Unit (ReLU) function is defined as follows [146]:

$$f(x) = max(0,x). \tag{2.25}$$

ReLU has the advantage of being very fast and, for classification problems, ReLU and its minor variants are “hard to beat”. However, ReLU might suffer from the “dying ReLU” issue. A combination of different circumstances can lead to “dead” ReLU neurons that are no longer able to become activated. To bypass this problem, “leaky ReLU” has been proposed and is defined as follows [146]:

$$f'(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases} \tag{2.26}$$

The swish function is an alternative, which outperforms ReLu in terms of accuracy. However, its computational costs are much higher. It is defined as follows [146]:

$$f(x) = x \cdot \sigma(x) = \frac{x}{1+e^{-x}}. \tag{2.27}$$

An additional non-zero parameter β might be included to create a range of different swish functions [146]:

$$f(x) = \beta x \cdot \sigma(\beta x) = \frac{\beta x}{1+e^{-\beta x}}. \tag{2.28}$$

This activation function is able to circumvent the vanishing gradient problem. The *softmax function* is often used as the activation function in the last layer of a neural network to normalize the output to a probability distribution over predicted output classes. It is represented as follows [128]:

$$\pi_\theta(s,a) = \frac{e_\theta^Q(s,a)}{\sum_{a'} e_\Theta^Q(s,a')}, \tag{2.29}$$

where $\pi_\theta(s,a)$ specifies the probability of selecting action a in state s. The softmax activation function has the following desirable properties [61]:

1. The sum of probabilities is always equal to 1.
2. The function is differentiable.
3. The function is an “adaptation of logistic regression to the multinomial case”.

Multilayer perceptrons

MLPs are the “simplest and most traditional architecture for deep learning models”. In this architecture, “the neurons in layer l_i are connected to every neuron in layer l_{i-1} with $i \in [1, L]$. These connections are modeled by the weights [...]. A general form of applying a non-linearity to an input time series X can be seen in the following equation” [61]:

$$A_{l_i} = f(w_{l_i} * X + b), \tag{2.30}$$

“with w_{l_i} being the set of weights with length and number of dimensions identical to X’s, b the bias term and A_{l_i} the activation of the neurons in layer l_i” [61]. Figure 2.16 shows the general structure of a MLP with an input layer, hidden layers and an output layer with each layer consisting of several neurons. The green neurons at the top represent the bias terms, while each link between neurons has an attached weight term.

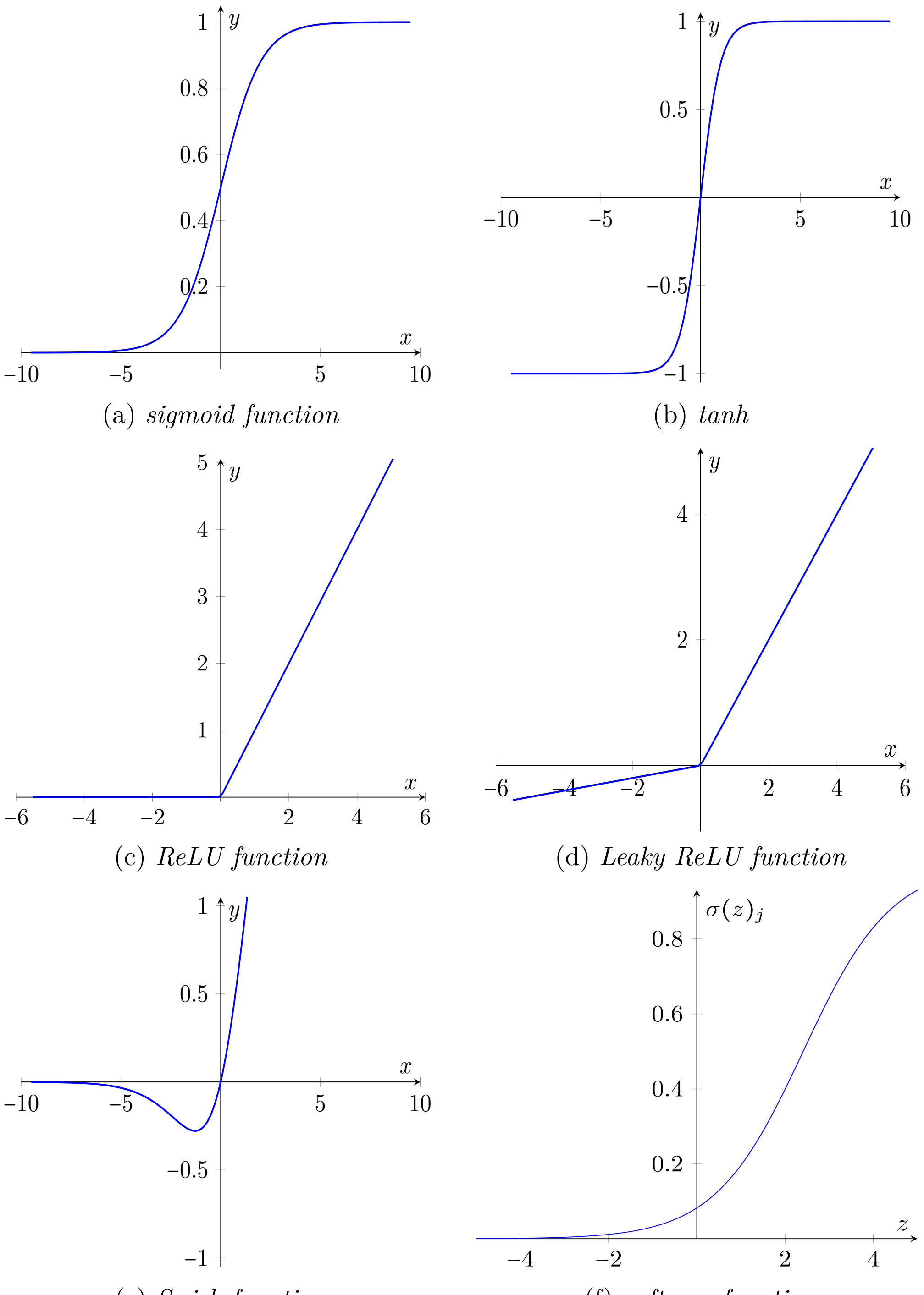
A disadvantage of using MLPs for TSC is that “each time stamp has its own weight and the temporal information [contained in the input data] is lost” during training [61].

Convolutional Neural Networks

CNNs are popular ANNs for image recognition, natural language processing and, more recently, also for TSC. Their core concept are convolutions, which can be interpreted as the process of sliding a 1-D filter over the time series for TSC. A convolution is applied for a time stamp t as follows [61]:

$$C_t = f(w * X_{t-l/2:t+l/2} + b) | \forall t \in [1, T], \tag{2.31}$$

where C is the result of a convolution “applied on a univariate time series X of length T with a filter w of length l, a bias parameter b and a final non-linear function f such as [ReLU]. The result of a convolution (one filter) on an input time series X can be considered as another univariate time series C that underwent a filtering process”. The sequential application of several filters enables CNNs to “learn filters that are invariant across the time dimension” [61].



(a) *sigmoid function*

(b) *tanh*

(c) *ReLU function*

(d) *Leaky ReLU function*

(e) *Swish function*

(f) *softmax function*

Figure 2.15.: Comparison of different activation functions for ANNs.

Residual Networks

Residual Networks (ResNets) have been described as the "deepest architecture[s]" for TSC. Their main concept is the residual bridge

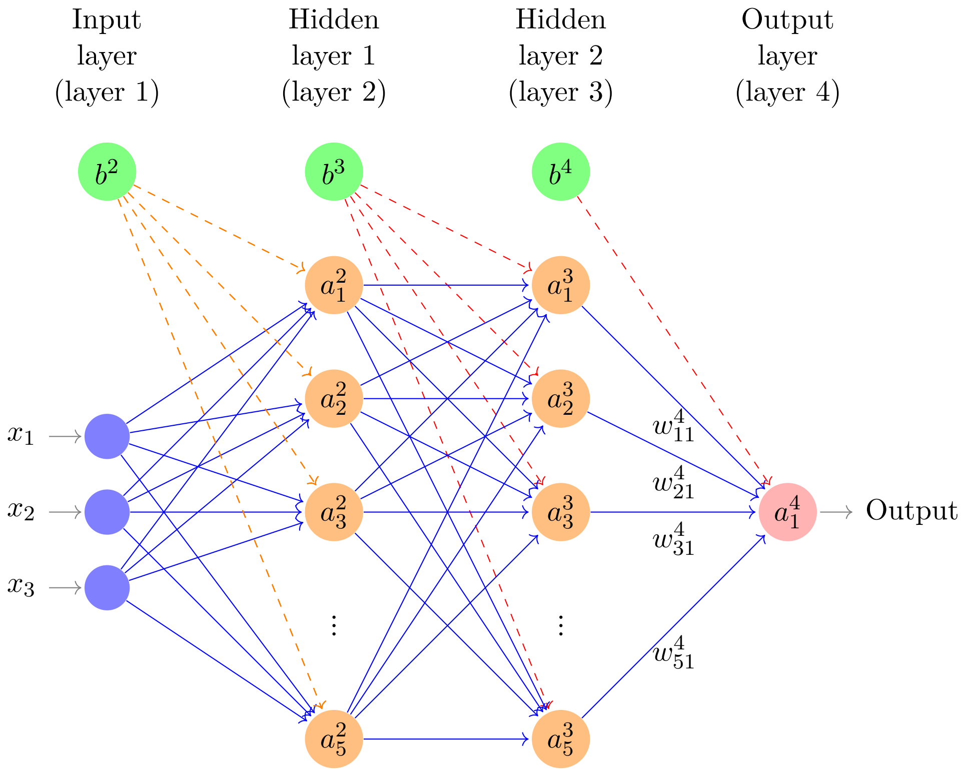


Figure 2.16.: General structure of a MLP showing input layer, hidden layers and output layers with each layer consisting of several neurons. The green neurons at the top represent the bias terms, while each link between neurons has an attached weight term.

connecting consecutive convolutional layers. Such a network contains “three residual blocks, [composed of three convolutions], followed by a [global average pooling] layer and a final softmax classifier” [61].

ROCKET

ROCKET transforms input time series with many kernels of “random length, weights, bias, dilation and padding”. The combination of ROCKET and Logistic Regression (see Section 2.7.4.1) results in an ANN with a single convolutional layer, where the transformed features are the input for a final trained softmax layer. The only adjustable hyperparameter for this model is the number of kernels, whereas more kernels result in higher classification accuracies but also in longer training times [83].

MiniROCKET

Even though ROCKET “achieves state-of-the-art accuracy with a fraction of the computational expense of most existing methods”, this method can still be further optimized with respect to efficiency and determinism. *MiniROCKET* reformulates ROCKET, making it “up to 75 times faster on larger datasets, [...] almost entirely deterministic” and, with additional computational expense, even "fully deterministic", while "maintaining essentially the same accuracy" as ROCKET [112]. MiniROCKET "transforms time series using convolutional kernels, and uses the transformed features to train a linear classifier" like ROCKET. However, unlike ROCKET, MiniROCKET uses a "small, fixed set of kernels [and] is almost entirely deterministic". MiniROCKET does not change two important aspects of ROCKET: dilation and proportion of positive values pooling. It is able to "massively reduce the time required for the transform" [112].

MultiRocket

Even though ROCKET and MiniROCKET are "two of the fastest methods for TSC", both are "somewhat less accurate" than HIVE-COTE and its variants, including HC-TDE (see Section 2.4.6.3). An alternative model called *MultiRocket* can significantly outperform both MiniROCKET and ROCKET with additional computational costs. This makes MultiRocket "the single most accurate method [...] while still being orders of magnitude faster than any algorithm of comparable accuracy other than its precursors" [113]. Like ROCKET and MiniROCKET, MultiRocket "transforms time series [with] convolutional kernels [and uses the transformed features to train] a linear classifier". MultiRocket makes use of several additional features, such as those present in the catch22 feature set (see Section 2.7.2.3) [98].

Transformers

Transformers is a ML architecture relying entirely on an attention mechanism to draw global dependencies between input and output. This model allows significantly more parallelization than traditional Recurrent Neural Networks (RNNs), LSTMs or gated RNNs approaches. It has been shown to achieve state-of-the-art results in machine translation [147]. Inspired by the success of Transformers, transformer based approach for (multivariate) TSC have been proposed [148]. An alternative transformer-based approach has also shown to achieve results "comparable to the state-of-the-art in photometric classification[s]" [149]. Transformers integrate an encoder-decoder structure using stacked self-attention and point-wise, fully connected layers [147].

2.7.4.7. Models not used in this work

There are some TSC models that have not been used in this work. This section aims to provide an overview of popular algorithms that have not been used in this work and states reasons for not including them.

2-D Artificial Neural Networks

This work does not perform TSC based on 2-D images as it aims to provide a comprehensive and efficient pipeline with a focus on mobile or wearable settings. The expected amount of additional memory, disk space and computational power needed for the creation of 2-D images from 1-D signals is not compatible with the overall goals of this thesis and prohibits the usage of such methods.

Adaptive Boosting

Adaptive Boosting (AdaBoost) is a statistical classification meta-algorithm of the Boosting family, which boosts the performance of a “weak” classifier by using it within an ensemble structure. The classifiers in the ensemble are added one at a time so that each subsequent classifier is trained on data which have been “hard” for the previous ensemble members. The ensemble construction through AdaBoost is equivalent to fitting an additive logistic regression model [150]. The *Rotation Forest* that has been presented for the first time in 2006, was compared with the standard implementations of Bagging, AdaBoost and Random Forest and “outperformed all three methods by a large margin” [150]. Hence, this work did not further explore the AdaBoost method.

Bayesian time series classification

Using *Bayesian inference* for TSC has been known to provide reasonable results for several decades with early works published in 1992 and 1995 showing the feasibility of this technique. In the past it has been successfully used to classify EEG signals [151]. More recently, a Bayesian CNN has been used to classify time series to detect marine gas discharges [152]. Previous works have also applied naive Bayesian classifiers, but these methods have been shown to “perform significantly worse” than sophisticated TSC algorithms, such as 1-NN DTW (see Section 2.7.4.2), SVM (see Section 2.7.4.5) or BOSS (see Section 2.4.3) [153]. Hence, this work does not further investigate the usage of (naive) Bayesian classifiers.

Dictionary based time series classification

WEASEL has been compared to BOSS, a randomised version of BOSS termed Randomised BOSS (RBOSS), BOSS vector space (BOSS-VS) and the meta ensemble HIVE-COTE on datasets from the UCR archive (see Section 2.4.3) [79]. This study reports that “WEASEL performed significantly better than all classifiers tested [but] it also comes in as the slowest classifier to build on average” [154]. Spatial pyramids in combination with BOSS have been found to be significantly more accurate than the original BOSS algorithm but are outperformed by HIVE-COTE (see Section 2.4.6) [155]. Hence, this work does not make use of any dictionary-based methods, as these methods have been shown to be less accurate and, in some cases, even slower than alternative approaches [83, 155, 96].

Gaussian Processes time series classification

ML models using Gaussian processes are a generalization of the Gaussian probability distribution. Like SVMs, they are a type of kernel model being a “traditional nonparametric tool for modeling”. It has been shown that Gaussian processes and infinitely wide deep neural networks are exactly equivalent to each other [156]. Furthermore, standard Gaussian Processes suffer from a cubic time complexity $\mathcal{O}(n^3)$ due to the inversion and determinant of the $n \times n$ kernel matrix. This limits its scalability and makes it unfeasible for large-scale datasets. With the help of approximation approaches, the complexity can be reduced to $\mathcal{O}(n \cdot m^2)$ with m inducing points. However, such approximations are still computationally impractical for real-time predictions [157]. Hence, this work does not further make use of Gaussian process classifiers.

Hidden Markov Models

Hidden Markov Models (HMMs) can model an observed sequence as probabilistically dependent upon a sequence of unobserved states. They have been found to be "computationally impractical for modeling long-range dependencies" [158]. As it is reasonable to assume that ANNs perform much better, this work does not further explore HMMs.

Recurrent Neural Networks

The training of traditional *RNNs* "has long been considered to be difficult" because issues such as *vanishing* and *exploding* gradients occur when backpropagating errors across many time steps. As *LSTM RNNs* use "carefully designed nodes with recurrent edges with fixed unit weight as a solution to the vanishing gradient problem" [158], this work does not further consider traditional RNNs for TSC.

LSTM RNNs have been proposed in the mid-1990s [107, 158] and have since "set records in accuracy on many tasks". In combination with *Bidirectional Recurrent Neural Networks (BRNNs)*, these networks have been applied successfully for phoneme classification and handwriting recognition [158]. RNNs, LSTM RNNs and gated recurrent neural networks have been established as state-of-the-art approaches in sequence modeling and transduction problems in the past. However, the more recent *Transformer* architecture allows for significantly more parallelization and can reach a new state-of-the-art in machine translation quality [147]. As Transformers have also shown superior performances in other fields, such as time series forecasting [159], this work does not investigate RNN-based models any further.

Relevance Vector Machines

RVMs are not included in further analyses in this work as their training involves the optimization of "a nonconvex function, and [RVM] training times can be longer than for a comparable SVM. For a model with M basis functions, RVM requires the inversion of a matrix of size $M \times M$, which in general requires $\mathcal{O}(M^3)$ computation" [77].

Random Forests

RFs are popular algorithms using decision trees to perform classifications. Random forests are a combination of trees that split the data according to a certain hierarchy. This work did not further explore RFs as Rotation Forests have been shown to outperform this algorithm “by a large margin” [150].

Rotation Forests

Rotation forests are tree-based ensembles that are “less well known and less frequently used” than alternative algorithms but are “significantly more accurate” on certain datasets [160]. However, a rotation forest is “relatively slow to build, particularly when the data has a large number of attributes”. Even though efforts have been described to mitigate this lack of efficiency [160], this work did not further explore Rotation forests due to very long tree building times.

Ensemble methods

Ensemble methods (see Section 2.4.6) use multiple weaker ML models and combine their outputs to obtain a better predictive performance. Ensemble methods, such as HIVE-COTE, TS-CHIEF, HC-CIF or HC-TDE have shown very promising results or even state-of-the-art performances. However, those methods have very high computational costs, which renders them infeasible for the datasets used in this work [96, 89, 97, 99].

Transfer learning

Transfer learning trains an ANN on a source dataset and then transfers the learned features to a second ANN meant to be trained on a target dataset. Transfer learning has been shown to perform better than traditional deep neural networks for tasks relying on computer vision [161]. However, more recently Transfer learning has also been applied to TSC tasks. *Timenet*, a multilayered Recurrent Neural Network (RNN), has been trained on 24 datasets from the *UCR Time Series Archive* (see Section 2.4 [162]. Another Transfer learning model trained on 85 datasets of the UCR archive “improve[d] or degrade[d] the models [performance] depending on the [dataset] used for transfer” [161]. A transfer learning method based on attributing sensor modality labels to a large amount of time series data has also been proposed [163]. In 2012, a publication presented a transfer learning algorithm “for adapting the [] propagation model to changing environments”. RF variations were modeled using GPs [164]. More recently, transfer learning has been used with *temporal enhanced ultrasound* data by learning from a RF ultrasound time series dataset and applying this newly gained knowledge to a B-Mode image dataset [165]. Additionally, the combination of a transfer learning model and a classic LSTM approach has also been used to distinguish between different devices [166]. Even more recently, “the applicability of pre-trained and retuned models to the RF domain” has been shown for several different tasks [167]. Even though transfer learning for Radio Frequency (RF) signals gained momentum and showed promising results in the last few years, this work does not make use of this technology as accessible pre-trained transfer learning models for US data were not available at the time of writing.

2.7.5. Model evaluation metrics

Being able to evaluate the performance is crucial in any ML pipeline. This section discusses common strategies and metrics, provides comparison and justifies the choices made for this work. Section 2.7.5.1 introduces the concept of *cross-validation (CV)*, while Section 2.7.5.2 provides background information about common metrics for binary classification evaluations.

2.7.5.1. Cross-validation

CV is an umbrella term for various similar model validation techniques for assessing how the results of a ML model will generalize to an independent dataset. It is mainly used in settings where the goal is prediction and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (*training dataset*) and a dataset of unknown data against which the model is tested (*validation dataset* or *testing dataset*). If we assume a study consisting of n subjects $S_1, S_2, \cdots, S_n$ with k measurements $M_1, M_2, \cdots, M_k$ for each subject, a common evaluation approach is leave-one-out cross-validation (LOOCV). Table 2.5 shows a sample database including four different subjects with various measurements for each subject.

Subject ID	**Measurement**
1	M_{11}
1	M_{12}
1	M_{13}
2	M_{21}
2	M_{22}
3	M_{31}
4	M_{41}
4	M_{42}

Table 2.5.: Sample database including four different subjects with various measurements for each subject.

Figure 2.17 shows a possible LOOCV approach for this particular example. This work deploys a similar approach for all TSC tasks detailed below.
This scenario groups the training and testing data according to association with the corresponding subject and creates four models in total. These models are then independently evaluated to obtain a metric of the general performance of the complete pipeline. Alternative approaches, such as grouping training and testing data according to certain data properties, are thinkable.

2.7.5.2. Binary classification evaluation

To evaluate the predictions of binary classification models, one can use several statistical metrics. Even though a suitable metric is crucial to evaluate ML models, “no widespread consensus has been reached on a

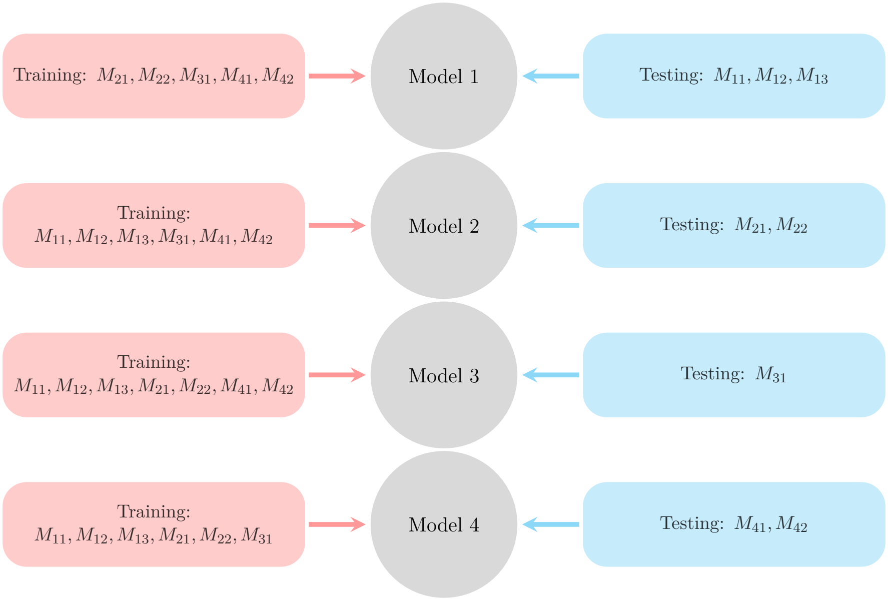

Training: $M_{21}, M_{22}, M_{31}, M_{41}, M_{42}$
Model 1
Testing: M_{11}, M_{12}, M_{13}
Training: $M_{11}, M_{12}, M_{13}, M_{31}, M_{41}, M_{42}$
Model 2
Testing: M_{21}, M_{22}
Training: $M_{11}, M_{12}, M_{13}, M_{21}, M_{22}, M_{41}, M_{42}$
Model 3
Testing: M_{31}
Training: $M_{11}, M_{12}, M_{13}, M_{21}, M_{22}, M_{31}$
Model 4
Testing: M_{41}, M_{42}

Figure 2.17.: Possible LOOCV approach for the database described in Table 2.5.

unified elective chosen measure yet. Accuracy and F_1-score computed on confusion matrices have been (and still are) among the most popular adopted metrics in binary classification tasks" [168].

Classwise performance metrics

While evaluating a ML model, every sample prediction belongs to one of the following four categories [168]:

1. *True positives (TP)*:
   "Actual **positives** that are correctly predicted **positives**".
2. *False negative (FN)*:
   "Actual **positives** that are wrongly predicted **negatives**".
3. *True negatives (TN)*:
   "Actual **negatives** that are correctly predicted **negatives**".
4. *False positives (FP)*:
   "Actual **negatives** that are wrongly predicted **positives**".

Based on those categories, a variety of performance measures can be derived. Due to various nomenclatures, the following list provides a brief overview [168]:

1. *Sensitivity*, *Recall* or *True positive rate*: $\frac{TP}{TP+FN}$
2. *Specificity* or *True negative rate*: $\frac{TN}{TN+FP}$
3. *Positive prediction value* or *Precision*: $\frac{TP}{TP+FP}$
4. *Negative prediction value*: $\frac{TN}{TN+FN}$
5. *False positive rate* or *Fallout*: $\frac{FP}{FP+TN}$
6. *False discovery rate*: $\frac{FP}{FP+TP}$

Confusion matrices

Table 2.6 shows the general structure of the standard *confusion matrix* describing the outcome and visualizing the performance of the ML model. Separately examining all four cells of a confusion matrix is time-consuming and not intuitive. Thus, additional metrics are needed for fast and straight-forward model performance comparisons.

	Predicted positive	**Predicted negative**
Actual positive	TP	FN
Actual negative	FP	TN

Table 2.6.: Standard confusion matrix.

Accuracy

The *accuracy* is "the ratio between correctly predicted instances and all the instances" [168]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.32}$$

The worst possible value is 0.0 and the best possible value is 1.0. Note that it has been argued that "accuracy fails in providing a fair estimate of the classifier performance in the class-unbalanced datasets" [168].

F_1-score

The F_1-score is "the harmonic mean of precision and recall" and is computed as follows [168]:

$$F_1\text{-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.33}$$

The worst possible value is 0.0 and the best possible value is 1.0.

Comparison of different performance metrics

Despite the frequent use of accuracy and F_1-score for ML model evaluations, a recent publication has shown that "these statistical measures can dangerously show overoptimistic inflated results, especially on imbalanced datasets". Matthews correlation coefficient (MCC) has been proposed as an alternative providing "a more informative and truthful score in evaluating binary classifications" [168]. MCC was originally developed in 1975 and re-proposed "in 2000 as a standard performance metric for [ML]". Despite its advantages, this work does not use the MCC and relies on the F_1 score instead as the former metric can face "situations - albeit extreme - where either MCC cannot be defined or it displays large fluctuations, due to imbalanced outcomes in the classification" [168].

Chapter 3

Classifying muscle contractions and muscle fatigue states

Assessment of specific muscle conditions is critical in many sports and rehabilitation settings. 1-D Sonomyography (SMG) relies on 1-D US signals to obtain information about deep soft tissue layers and enables cost-effective and portable solutions. This chapter illustrates the use of different ML approaches for two important scenarios: The classification of contracted muscles versus non-contracted muscles and the classification of relaxed muscles versus fatigued muscles. To this end, the ML models presented include muscle contraction signals from eight volunteers and muscle fatigue signals from 21 volunteers obtained with 1-D SMG. To mimic real-world scenarios as closely as possible, the experiments presented do not rely on carefully selected and unique signals from a restricted body region. This chapter describes ML models based on different evaluation schemes including either all signals, only signals from a specific arm, only signals from a specific gender, or only signals from a single individual to illustrate the respective effects on the results. In addition, it evaluates ML model performances based on a variety of different data types or features extracted from the acquired input signals.
Partial results of the work presented in this chapter have been published in [2].

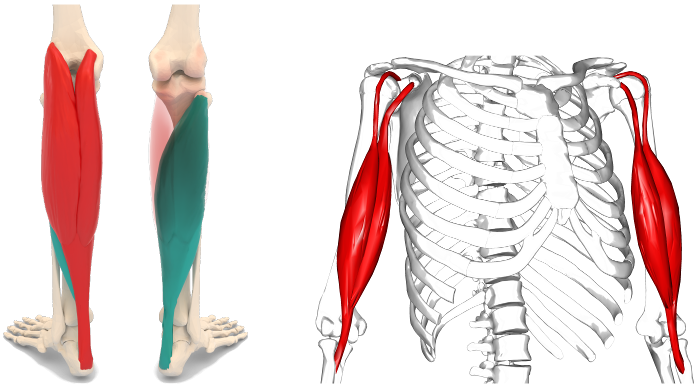
Contents

3.1. Introduction

Section 3.1.1 motivates this chapter, while Section 3.1.2 explains underlying muscle physiology.

3.1.1. Motivation

The assessment of different *muscle contraction* and *muscle fatigue* states is crucial in many fitness and physical rehabilitation scenarios, such as for the assessment of therapeutic measures. Potential applications based on the methods described in this chapter might include smart and mobile low-cost devices quantifying muscle state changes to track a person's fitness or rehabilitation level. To this end, mobility and wearability aspects are of particular importance to increase the device's suitability for daily use. This chapter focuses on classifying 1-D US signals with ML approaches. One objective of this work is to address drawbacks of alternative approaches (see Section 3.2) by using signals from deeper soft tissue layers that have been acquired with a wearable device. This work does not take technical and practical considerations into account, which remains to be done in future works.



(a) Gastrocnemius muscle.

(b) Biceps brachii muscle.

Figure 3.1.: Gastrocnemius and biceps brachii muscles (reprinted with permission from [169] (CC BY-SA 2.1 JP).

3.1.2. Muscle physiology

Figure 3.1 illustrates the examined muscles or muscle groups. The gastrocnemius muscle is sketched in Figure 3.1 (a) and the biceps brachii muscle is sketched in Figure 3.1 (b).

3.1.2.1. Physiology of muscle contractions

The study of muscle contractions has been conducted for centuries but in recent years insights of the underlying physiology of muscle contractions have steadily increased. In the early 20th century, the *Hill's elastic body theory*, developed by the British physiologist Archibald Vivian Hill (1886-1977), considered a stimulated muscle to be a “new elastic body”. However, this theory was later disproved. In 1954, two publications from the British molecular biologist Hugh Huxley (1924-2013) and the British physiologist and biophysicist Andrew Fielding Huxley (1917-2012) independently proposed that muscle contraction occurred “by the relative sliding of two sets of filaments, actin and myosin. In 1957, Andrew Huxley proposed how this relative sliding might occur, and provided a mathematical framework for what is now known as the *crossbridge theory of muscle contraction*”. However, this theory failed to properly predict certain properties of actively stretched muscles as the predicted “forces and energy consumption were much too big compared to experimental results and residual force enhancement could not be predicted conceptually”. To address these theoretical flaws, a publication from 2015 proposed a sophisticated muscle contraction theory involving the three filaments *actin*, *myosin* and *titin* [170]. Future research will show whether this theory can explain all phenomena occurring in real physical experiments or whether it must be amended further to address potential shortcomings.

3.1.2.2. Physiology of muscle fatigue

Muscle fatigue is defined as an exercise-induced reduction in *maximal voluntary contraction (MVC)* [171]. It can originate at different levels of the motor pathway and is usually divided into central and peripheral components. *Peripheral fatigue* is produced by changes at the neuromuscular junction. *Central fatigue* originates at the central nervous system and decreases the intensity of motor neuron spikes directed at the muscle in question. Muscle force production involves a sequence of events, extending from cortical excitation to motor unit activation to excitation-contraction coupling and ultimately leading to muscle activation. Changes in the nervous, ion, vascular and energy systems, impair force generation and contribute to the development of muscle fatigue. Metabolic factors and fatigue reactants also affect muscle fatigue [172]. This work focuses on muscle fatigue during or after intense exercise.

3.2. State of the Art

3.2.1. Electromyography

EMG is a technique created for evaluating and recording the electrical activity produced by skeletal muscles. There are two kinds of EMG: Surface Electromyography (sEMG) and Intramuscular Electromyography (iEMG). The former assesses muscle functions by recording muscle activity from the surface above the muscle on the skin, while the latter typically relies on a needle electrode that is inserted invasively directly into the muscle tissue through the skin. Section 3.2.1.1 and Section 3.2.1.2 describe both approaches in more detail.

3.2.1.1. Surface Electromyography

sEMG works by placing electrodes directly over the muscle locations on the skin surface. These electrodes are a non-invasive technique and easy to apply. However, they can only be applied on superficial muscles that are large enough to support electrode mounting on the skin surface. Additionally, crosstalk of signals stemming from various muscles is a challenge for smaller muscles within a complex mechanical arrangement, such as the forearm [173]. sEMG is applied in many fields, such as motor control of human movement, myoelectric control of prosthetic and orthotic devices, in rehabilitation scenarios, in gesture recognition interfaces and in gesture control and motion control devices [125, 173]. Figure 3.2 shows a sample sEMG signal.

3.2.1.2. Intramuscular Electromyography

iEMG either relies on needles or fine wire electrodes that are inserted through the skin directly into the muscle. As the needles need to be positioned correctly, this technique requires trained professionals in comparison to sEMG (see Section 3.2.1.1). Although iEMG needles are ideal

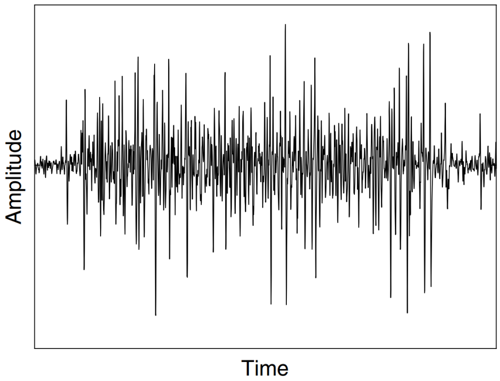


Figure 3.2.: Sample sEMG signal (taken from [125]).

for recording deep muscle activities, their correct placement "requires a detailed knowledge of musculoskeletal anatomy". Even though a recent review published in 2021 found iEMG to be a useful tool to analyze a variety of different muscle activations in different activities [173], the invasiveness of needles and the associated pain are major disadvantages of intramuscular electrodes.

3.2.2. Inertial Measurement Units

Inertial Measurement Units (IMUs) are electronic devices that measure and report a body's specific force, angular rate and sometimes the orientation of the body, using a combination of accelerometers, gyroscopes or magnetometers. Besides navigational purposes, IMUs serve in almost all smartphones and tablets as orientation sensors. Fitness trackers and other wearables may also include IMUs. These sensors are the most widely used wearable sensors for *gait analysis*. However, IMUs do not measure electrical activity in muscles, which gives them an inherent disadvantage over alternative methods, such as EMG (see Section 3.2.1) [174]. They do not provide any information of deeper soft tissue layers either.

3.2.3. Force Sensitive Resistors

Force Sensitive Resistors (FSRs) are polymer thick film devices that change their electric resistance according to the force applied to their active surface. "The resistance is inversely proportional to the applied load". If unloaded and unbent, the stand-off resistances of the FSRs are typically very high and decrease if loaded and bent. FSRs require only a simple interface, consume little power, are unobtrusive and lightweight. In a publication from 2006, it has already been shown that FSRs are a viable alternative method to detect muscle activities, while being "ideal for wearable applications" [175]. More recent work presented a dual-channel, non-invasive *force myography sensor* to extract muscle contraction information for controlling hand prostheses. This sensor was prepared using a pair of FSRs mounted inside a rigid base for sensing the force exerted by contracting muscles through *polydimethylsiloxane* couplers. Experiments revealed that this sensor may provide an alternative to EMG (see Section 3.2.1) [176].

3.2.4. Textile capacitive and pressure mapping sensors

Textile capacitive sensors consist of a three-layer structure forming a capacitance with a pressure sensing non-conducting dielectric. These sensors have been proposed in a publication from 2006 and were meant as an alternative to FSRs (see Section 3.2.3), which “have a certain stiffness and could therefore affect the comfort of the wearer negatively” [177]. More recent related work presented a wearable textile sensor system for monitoring muscle activities by leveraging surface pressure changes between skin and compression garment. A comparison with arm-worn EMG (see Section 3.2.1) shows that this “approach is comparable on the signal quality level” [178].

3.2.5. Epidermal Electronics

Epidermal Electronics are integrated electronics that are ultrathin, soft and lightweight and can be mounted to the epidermis based on van der Waals interactions. A study published in 2020 showed that pencil drawings of a variety of bioelectronic devices on commonly used office-copy papers with commercial pencils work as conductive traces and sensing electrodes. Using this approach, a variety of different sensors can be created, including temperature sensors, electrophysiological sensors, electrochemical sweat sensors, joule-heating elements and ambient humidity energy harvesters. With the help of those sensors, a variety of vital biophysical information can be obtained. This information includes ”skin temperatures, ECG, EMG, alpha, beta and theta rhythms, instantaneous heart rates and respiratory rates in a real-time, continuous and high-fidelity manner”. EMG signals can be collected from the forearm of a human using pencil-paper on-skin electrophysiological sensors [179]. Figure 3.3 shows several aspects of pencil-paper on-skin sensors. These include a long-term ECG recording (F), a dynamic amplitude response of the R peaks in ECG signals (G), respiratory rates determined by R peak amplitude variations (H), comparisons of instantaneous heart rates determined from the R-to-R peak intervals in the ECG signals (I) and EMG signals recorded from the forearm of a human volunteer (J).

3.2.6. B-Mode Ultrasound

B-Mode US is often used to study human skeletal muscle anatomies (“e.g. muscle belly length, thickness and cross-sectional area”). Evidence from studies shows that B-Mode US is a reliable method to determine fascicle architectures during movements involving voluntary muscle contractions. This architecture is of particular importance “because of its relation with sarcomere length and, hence, its (indirect) relation with muscle force-producing capability and energetics” [180]. B-Mode US is a

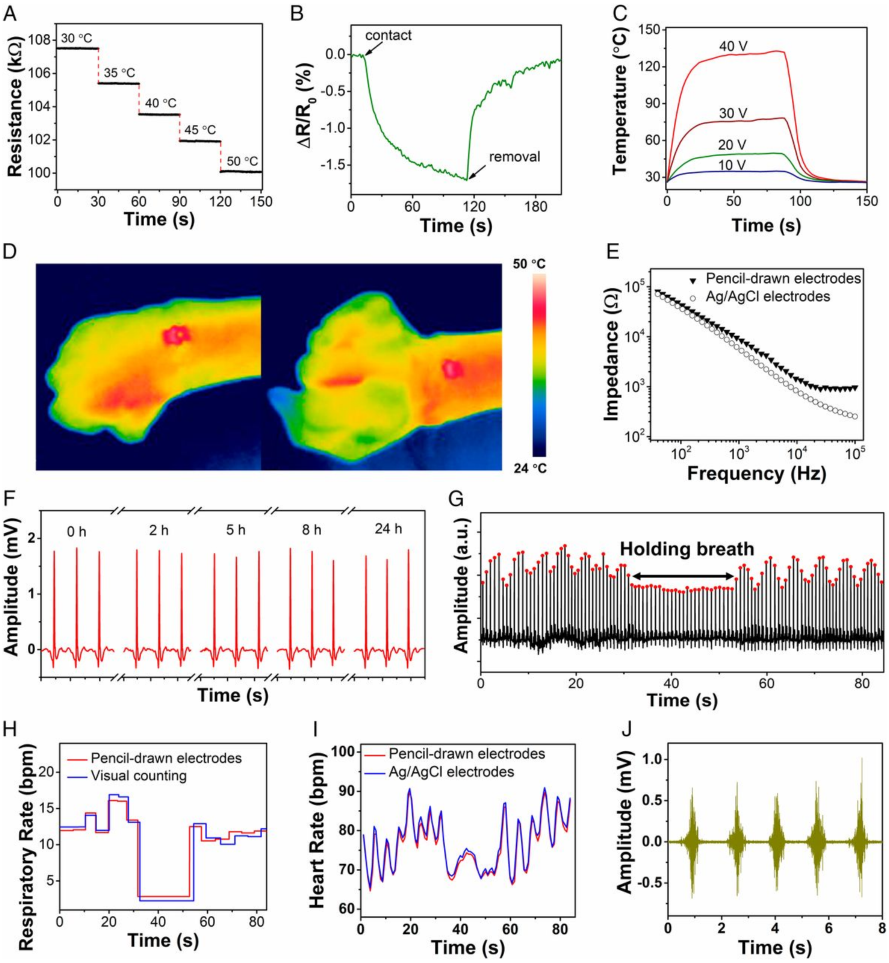


Figure 3.3.: Several technological aspects of epidermal electronics. Subfigure F shows a long-term ECG recording from the chest of a human volunteer using the pencil-paper on-skin electrophysiological sensor (taken from [179]).

non-invasive technique producing very intuitive images that are easy to interpret for medical practitioners.

3.2.7. A-Mode Ultrasound

A-Mode US only requires a single element US transducer to transmit and receive US waves or their respective echoes. This technique has several advantages over B-Mode US (see Section 2.1.4.3). However, sophisticated DSP or ML algorithms are needed to process the A-scans acquired with A-Mode US. There are several applications for 1-D SMG. Section 3.2.7.1 discusses solutions for *muscle contractions and prosthesis control*, while Section 3.2.7.2 focuses on applications for *muscle fatigue*.

3.2.7.1. Muscle Contractions and Prosthesis Control

Early work, published in 2008, introduced *1-D SMG* relying on A-scans acquired with a 10 MHz single-element US transducer to detect dynamic thickness changes in skeletal muscle during contraction. The results demonstrated that 1-D SMG "can be reliably performed and that it has the potential for skeletal muscle assessment and prosthesis control" [181]. Work published in 2015 introduced a wearable *US radial muscle activity detection system*, which recognized "forearm muscle activities of amputee subjects to control a dexterous prosthetic hand". The system consisted of control electronics to capture and record the US echo signals in real-time and two wearable bands embedded with eight 5 MHz single-element US transducers. That work also found that conventional sEMG (see Section 3.2.1.1) "cannot reliably identify the deeper muscle activations in the forearm due to crosstalk and signal attenuation. It also suffers from signal degradation due to muscle fatigue and non-linearity" [182]. Subsequent work published in 2016 presented and validated a finger gesture recognition method that acquired A-scans with four 5 MHz single-element US transducers, which were integrated in an armband [183]. Further work, published in 2018, presented the design, simulations, fabrication and evaluation of single and dual frequency US transducers for dexterous gesture recognition. To this end, four single-element US 5 MHz transducers were placed on the brachioradialis muscle, flexor carpi radialis muscle, flexor digitorum superficialis muscle and flexor carpi ulnaris muscle respectively. That study found dual frequency US transducers to yield a better performance than single frequency US transducers [184]. Work published in 2018 found that it is evident that sEMG based human-machine interfaces show "inherent difficulty in predicting dexterous musculoskeletal movements such as finger motions" and presented an alternative based on four single-element 5 MHz US transducers for a lightweight device, that was adopted to evaluate the performance of finger motion recognition. The outcomes confirmed the feasibility of A-Mode US based wearable human-machine interfaces. All previous work regarding 1-D SMG were analyzed offline and that work was the first to further evaluate its feasibility by performing an online evaluation. To this end, the authors deployed LDA and SVM algorithms for signal classifications [185]. Another work published in 2018 applied sEMG and A-Mode US to detect muscle deformation and motor intent.

To this end, experiments combining US, obtained with a 5 MHz single-element US transducer, and sEMG were conducted and compared to US only and sEMG only experiments. By applying the SVM algorithm, elbow angle and torque could be reconstructed from A-scans[186]. In 2019, sEMG and A-Mode US were compared for gesture recognition and isometric muscle contraction force estimation tasks. Five-channel A-Mode US transducers with a 5 MHz central frequency and five bi-polar sEMG were utilized and it was found that A-Mode US outperformed sEMG on gesture recognition accuracy, robustness and discrete force estimation accuracy, while sEMG was superior on continuous force estimation accuracy and ease of use in force estimation [187]. In the same year, a customized wearable forearm armband consisting of eight single-element US transducers with a central frequency of 5 MHz was used for muscle contraction detection and yielded superior performances on excitation pulse, detection depth and axial resolution. Moreover, in vivo muscle deformation detection and virtual prosthesis control experiments demonstrated its ability for rehabilitation applications, such as stroke rehabilitation and prosthesis control [188]. Another study, published in 2019, proposed a portable hybrid sEMG / A-Mode US system for human-machine interfaces, consisting of a wearable armband featuring sEMG electrodes and four 5 MHz single-element US transducers. The proposed system met the requirements for simultaneously acquiring high-quality sEMG and A-Mode US signals from the same muscle, for relatively good wearability and for demonstrating both the electrophysiological and the morphological information of the muscle. The hand gesture recognition experiment based on that proposed system verified the benefit of combining sEMG and A-Mode US together [189]. Four-channel A-Mode US transducers with a central frequency of 5 MHz were used in another study from 2019 to achieve gesture recognition and muscle contraction force estimation. The authors demonstrated a “more robust gesture recognition performance during force variation and comparable force estimation precision” in comparison to sEMG [190]. The same year saw the publication of a study examining whether the acoustic non-linearity parameter (B/A) can be used to partially represent the contraction state of skeletal muscles. This parameter “is defined as the ratio of the coefficients of quadratic terms to those of linear terms in the Taylor expansion of the state equation” [191]:

$$B/A = 2\rho_0 c_0 (\delta c/\delta p)_{0,s}, \tag{3.1}$$

where ρ_0 is the density, c_0 the velocity, p the static pressure and s the entropy. In that work, several experiments with A-Mode US were conducted to prove that the B/A value has “the potential to dynamically represent skeletal muscles” [191]. In 2020, a study introduced single-element wearable A-Mode US sensors and a method to measure skeletal muscle contractile parameters with them. The developed sensor was “employed to monitor the contractions of [the gastrocnemius] muscle of a human subject”, whose contractions were evoked by electrical muscle stimulations. The authors concluded that their approach “could be a valuable tool for inexpensive, non-invasive and continuous monitoring of the skeletal muscle contractile properties” [58].

3.2.7.2. Muscle Fatigue

Published literature about quantifying *muscle fatigue* with A-Mode US is much less common in comparison to published literature about *muscle contractions*. A study published in 2017 used a custom-made single-element transducer with a central frequency of 2.25 MHz that was integrated into an armband to continuously monitor thickness changes of the biceps brachii muscle during fatigue inducing exercises. The authors concluded that “those custom-made single-element transducers [could] effectively extract muscle information to assess muscle fatigue” [192].

3.2.8. Comparison

Table 3.1 summarizes the disadvantages and advantages of the methods described above.

Technique	Non-invasiveness	Wearability	Ability to obtain signals from deeper muscle layers
Intramuscular Electromyography	- -	- -	+ +
Surface Electromyography	+ +	+ +	- -
Inertial Measurement Units	+ +	+ +	- -
Force Sensitive Resistors	+ +	+ : + +	- -
Textile capacitive and pressure mapping sensors	+ +	+ +	- -
Epidermal Electronics	+ +	+ +	- -
B-Mode Ultrasound	+ +	- -	+ +
A-Mode Ultrasound	+ +	+ : + +	+ +

Table 3.1.: Summary comparing all methods mentioned above.

1-D SMG is a suitable technology for non-invasive and wearable solutions to obtain deep muscle layer signals for muscle contraction and muscle fatigue state classification tasks. However, technical considerations remain as A-Mode US necessarily requires ultrasound gel or other coupling materials to avoid large impedance mismatches, which has a negative effect on the wearability of potential systems. See Section 6.4.1 for a thorough discussion of this challenge and its implications.

3.3. Materials and Methods

This section describes the experimental designs and the acquired databases for both the muscle contraction states classification and the muscle fatigue states classification scenarios in Section 3.3.1 and Section 3.3.2 respectively. Section 3.3.3 illustrates details of the different signal types used for classification, while Section 3.3.4 provides details of the approaches used to annotate the data. Section 3.3.5 provides an overview of the evaluation schemes used to quantify the performances of different signal types and ML models.

3.3.1. Experimental design

The signals for both muscle contraction states classification and muscle fatigue states classification were acquired with the single-element US transducer *Olympus Panametrics™ V542-SM* with a center frequency of 2.25 MHz. All subjects were recruited from the *Fraunhofer Institute for Biomedical Engineering (IBMT)* in Germany and were aged between 20 and 50 years. Section 3.3.2 describes the complete database in detail.

3.3.1.1. Custom hardware and software

Mobile, low-cost and low-power lithium-ion battery powered US electronics were developed by the *Fraunhofer Institute for Biomedical Engineering (IBMT)* to acquire 1-D US signals for this work and other projects. These electronics communicate with external devices via the wireless module *Espressif ESP-12F*. A transmission pulser generates rectangular, bipolar burst signals and the received US echoes are sampled with 40 MHz using eight parallel electrical channels of the transceiver *Maxim MAX2082*. A low cost and low power *Xilinx Artix-7 Field Programmable Gate Array (FPGA)* controls all other components and is in charge of data management. The system has been designed for US transducer frequencies ranging from 0.5 to 5 MHz. The modular design of the system features a pluggable power board and main board as *printed circuit boards* with the dimensions 80 x 31 x 16 mm. Figure 3.4 shows the core US acquisition electronics developed at the *Fraunhofer Institute for Biomedical Engineering*.

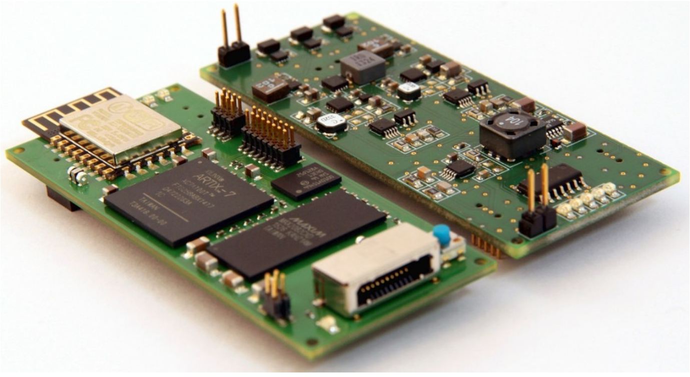


Figure 3.4.: Core US acquisition electronics (courtesy of Fraunhofer Institute for Biomedical Engineering).

A custom app for the mobile operating system *Android*, created with the open-source framework *Xamarin* for *.NET* and *C#*, allows the storage, visualization and transmission of the acquired signals.

3.3.1.2. Muscle contraction state classifications

Eight healthy subjects gave their consent to participate in the muscle contraction states classification experiment. They performed squats to allow the acquisition of signals belonging to contracted and non-contracted muscle states, respectively. The subjects fixated the single-element US transducer above their Gastrocnemius calf muscle, which is located on the back of the lower leg, without receiving any further instructions on any specifically distinct calf muscle locations. Even though this approach might have resulted in the acquisition of signals suffering from avoidable disturbances, such as interference from neighboring muscles or muscle groups, it mimicked real-life scenarios adequately. Requiring users to investigate possible best fitting muscle positions or even using B-Mode US first to locate a suitable muscle position, as has been done before [181], would significantly affect the usability and user acceptance of the system. Figure 3.5 shows the experimental setup for the muscle contraction signal acquisitions. The US transducer was attached to the leg via a supporting holding and a stretch band. The device to the right was an accelerometer, which acquired additional data concerning the spatial orientation of the leg.

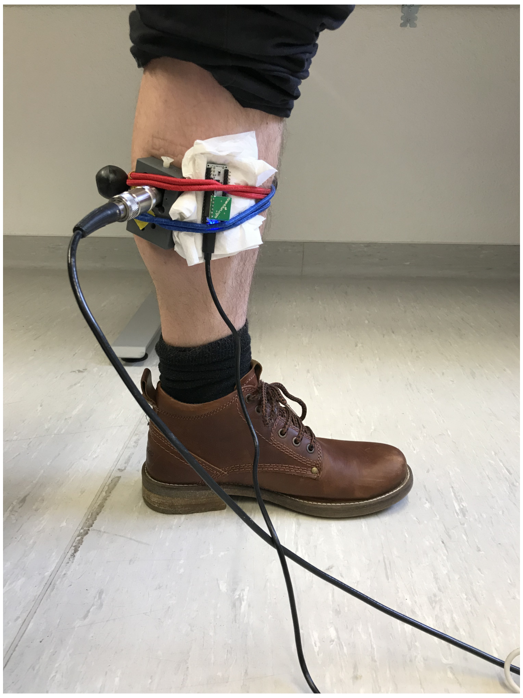


Figure 3.5.: The US transducer was attached to the leg via a supporting holding and a stretch band. The device to the right was an accelerometer, which acquired additional data concerning the spatial orientation of the leg.

3.3.1.3. Muscle fatigue state classifications

21 healthy subjects gave their consent to participate in muscle fatigue classification experiments, in which they were asked to lift weights chosen according to their subjectively perceived fitness level. All participants were instructed to lift the weights as long as possible to induce muscle fatigue. As described in Section 3.3.1.2, the subjects of the muscle fatigue study also did not put any emphasis on obtaining particularly distinct signals by choosing the region of interest on the skin surface particularly carefully. Instead, the volunteers put the US transducer on any fitting area above the biceps brachii muscle to simulate real-life scenarios. Figure 3.6 shows the experimental

setup for the acquisition of muscle fatigue state signals. This figure shows a study subject lifting a weight, while a single-element US transducer is attached to the body surface via a stretch armband. The signals were acquired with custom-made acquisition hardware, which transferred them wirelessly to a mobile device. After acquisition, the signals were transferred to a computer performing offline classifications.

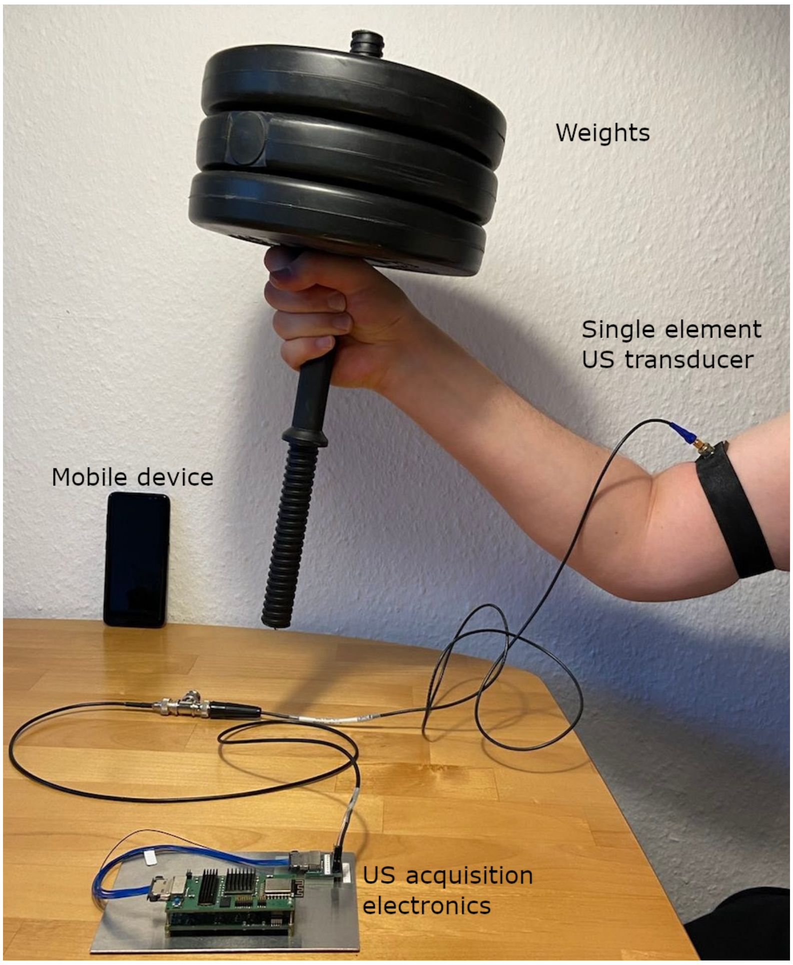


Figure 3.6.: Experimental setup showing a subject lifting a weight, while a single-element ultrasound transducer is attached to the body surface via a stretch armband (taken from [2]).

3.3.2. Database

3.3.2.1. Muscle contraction signals

For the muscle contraction signal acquisition, eight healthy volunteers agreed to participate in the experimental study described in Section 3.3.1.2. Seven subjects were male and one was female. The complete database is available online [7]. Table 3.2 provides a summary of the sex of each subject, the amount of acquired A-scans for each subject, the duration of the acquired dataset and the average amount of A-scans acquired per second for each dataset. Furthermore, the table notes whether the position of the US transducer on the Gastrocnemius calf muscle was unique (i.e. different from any other previously chosen location) or not (i.e. exactly the same as a location on a previously chosen location). This attribute is relevant because the results presented in Section 3.4 only include signals stemming from the same (i.e. non-unique) US transducer positions.

Dataset ID	**Subject ID**	**Sex**	**Unique transducer position**	**# A-scans**	**Duration (s)**	**A-scans / s**
1	01	male	no	3,000	54.834	54.71
2	01	male	no	3,000	59.351	50.55
3	02	male	no	1,000	18.077	55.32
4	02	male	no	1,000	19.209	52.06
5	01	male	yes	6,000	106.819	56.17
6	01	male	yes	50,000	1,472.675	33.95
7	03	female	yes	10,000	205.931	48.56
8	03	female	yes	8,872	157.950	56.17
9	04	male	yes	10,000	205.560	48.65
10	04	male	yes	10,000	256.566	38.98
11	05	male	yes	10,000	244.964	40.82
12	06	male	yes	10,000	220.608	45.33
13	07	male	no	10,000	170.281	58.73
14	07	male	no	10,000	167.926	59.55
15	01	male	yes	10,000	166.762	59.97
16	01	male	yes	10,000	165.754	60.33
17	01	male	yes	10,000	166.149	60.19
18	08	male	no	10,000	210.185	47.58
19	08	male	no	10,000	210.599	47.48
20	08	male	yes	10,000	183.689	54.44
21	08	male	yes	10,000	222.795	44.88

Table 3.2.: Muscle contraction signals database.

3.3.2.2. Muscle fatigue signals

For the muscle fatigue signal acquisition, 21 healthy volunteers agreed to participate in the experimental study described in Section 3.3.1.3. Fourteen subjects were male and seven were female. The complete database is available online [8, 9]. Table 3.3 and Table 3.4 provides summaries of the sex of each subject, the amount of datasets of each subject, the amount of A-scans of each dataset, the duration of all datasets and the maximum lifted weight for each subject of the first and second study, respectively. The first study enables the creation of ML models yielding robust and reliable results for signals stemming from a large variety of subjects, while the second study enables the creation of robust and reliable ML models for signals stemming from a single male subject only. Note that subject 09 from the first study is identical to subject 09 from the second study.

Subject ID	Sex	# Datasets	# A-scans	Total duration of all datasets (s)	Maximum lifted weight (kg)
01	female	4	1,390	1,040.74	5.0
02	female	3	1,043	865.00	2.5
03	male	2	696	925.99	2.5
04	male	2	685	1083.18	2.5
05	male	2	695	241.22	7.5
06	male	4	1,386	1553.51	7.5
07	female	4	1,367	770.52	5.0
08	male	3	1,044	496.66	5.0
09	male	10	3,453	1,808.52	7.5
10	female	3	1,044	449.05	5.0
11	female	2	696	183.9	5.0
12	male	2	695	243.51	5.0
13	male	3	1,035	840.86	7.5
14	female	2	695	393.56	5.0
15	male	2	666	740.79	5.0
16	male	2	672	586.41	5.0
17	male	1	348	215.00	7.5
18	male	3	1,035	800.72	5.0
19	male	1	342	149.30	7.5
20	male	1	345	137.24	7.5
21	female	1	345	338.07	2.5

Table 3.3.: Muscle fatigue signals database for all subjects [study 1].

Subject ID	Sex	# Datasets	# A-scans	Total duration of all datasets (s)	Maximum lifted weight (kg)
09	male	42	13,160	4,879.33	7.5

Table 3.4.: Muscle fatigue signals database for a single subject [study 2].

3.3.3. Signal types

Figure 3.7 illustrates the difference between a raw signal taken from the muscle fatigue database and its truncated version. The upper figure displays the complete signal, while the lower figure shows the same signal truncated to the interval ranging from sample index 500 to sample index 2000. Figure 3.8 shows raw 1-D US A-scans and the transformed signals of a relaxed and fatigue muscle stemming from the same subject. The signals in this figure are either not processed at all, filtered with a Butterworth band pass filter (see Section 2.3.2.4) or transformed with the Fourier transform (see Section 2.3.2.1), Wavelet transform (see Section 2.3.2.2) or Hilbert transform (see Section 2.3.2.3). Statistical, spectral, temporal features or a combination thereof are also taken into consideration as input data (see Section 2.3.3). The first and last two seconds of each dataset are ignored to account for noise that might have stemmed from lifting or depositing the weight at the beginning or end of each weight lift.

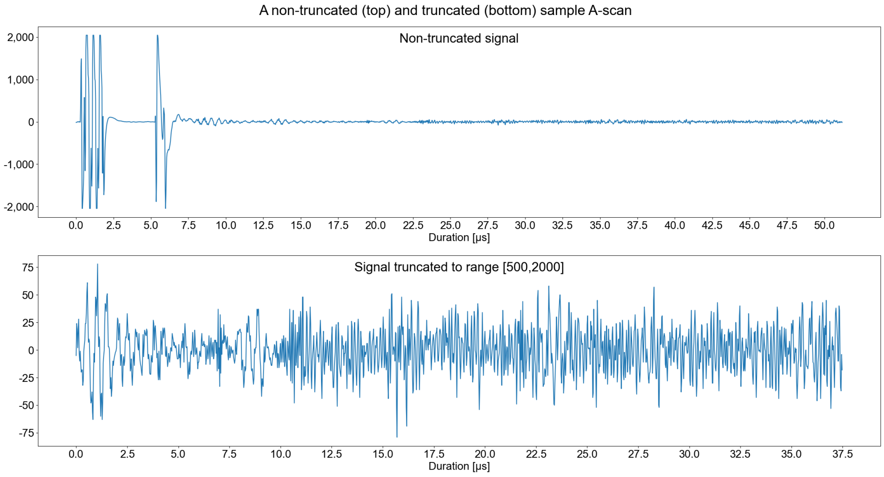


Figure 3.7.: Comparison of a raw 1-D US muscle fatigue A-scan and its truncated version. The upper figure displays the complete signal, while the lower figure shows the same signal truncated to the interval ranging from sample index 500 to sample index 2000.

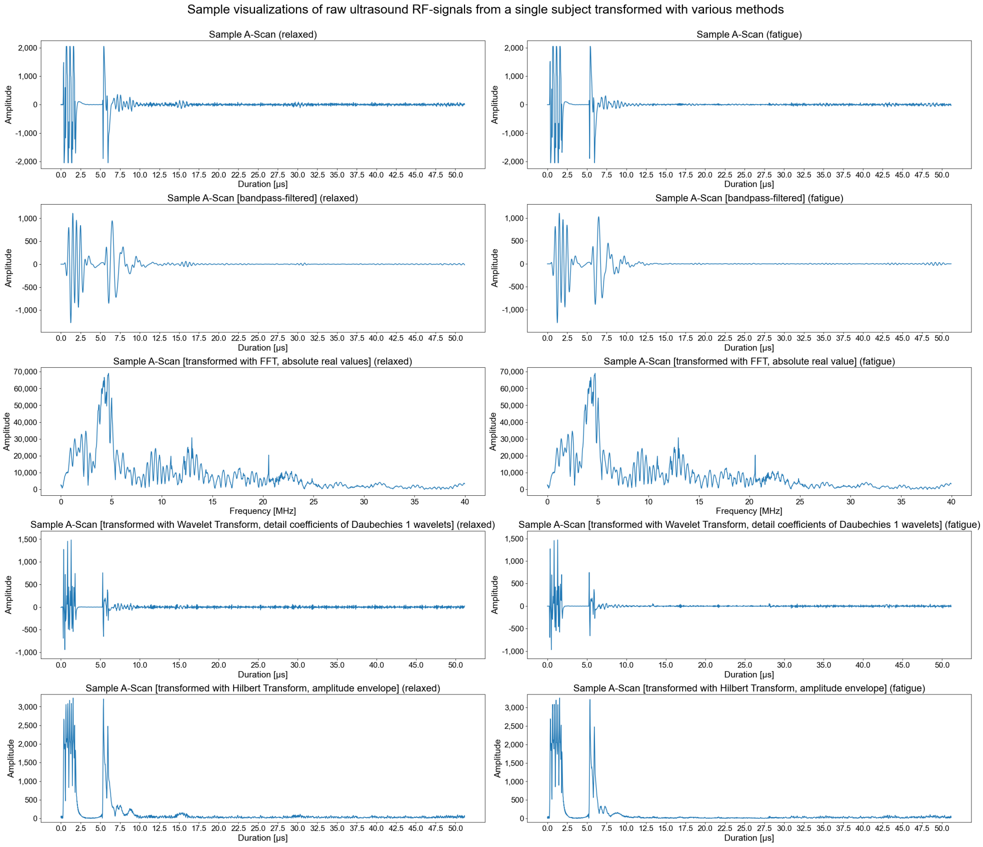


Figure 3.8.: Raw 1-D US signals and respective transformations of a relaxed and fatigue muscle stemming from the same subject (taken from [2]).

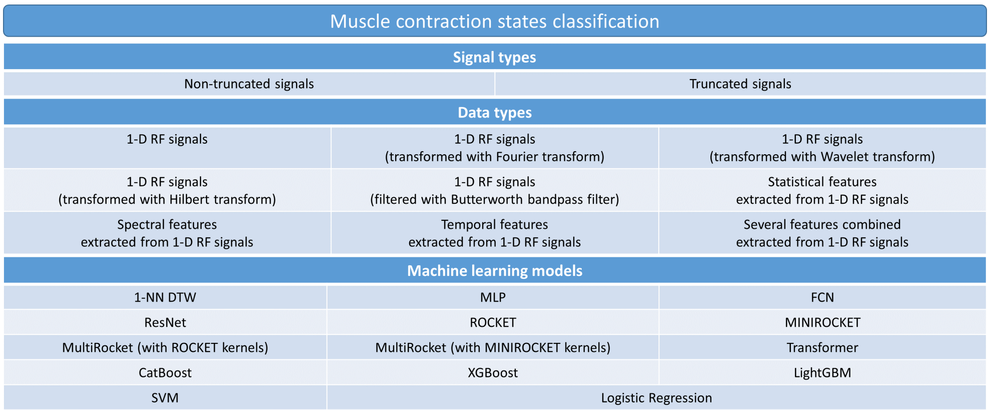
3.3.4. Annotations

All subjects in the muscle contraction state classification study annotated the signals by pushing a button every time they performed a squat. To this end, a custom-designed software tool tracked each push of the button. For the muscle fatigue state classification study, signals stemming from the first 10 seconds of each dataset represented the ”normal” category, while all signals stemming from the last 10 seconds of each dataset represented the ”fatigue” category.

3.3.5. Evaluation schemes

3.3.5.1. Muscle contraction evaluation scheme

For muscle contraction state classifications, this work used only signals stemming from the *same person* and the *same transducer position*. These signals represented 22.55 % of all available A-scans, 38.1 % of all acquired datasets and entailed signals from 37.5 % of all subjects. The inclusion of omitted signals did not result in models performing better than random guessing, which is most probably due to an insufficient size and diversity of the database. Figure 3.9 shows a hierarchical diagram illustrating all computed input signal combinations for the muscle contraction state classification signals.



Muscle contraction states classification

Signal types

Non-truncated signals	Truncated signals

Data types

1-D RF signals	1-D RF signals (transformed with Fourier transform)	1-D RF signals (transformed with Wavelet transform)
1-D RF signals (transformed with Hilbert transform)	1-D RF signals (filtered with Butterworth bandpass filter)	Statistical features extracted from 1-D RF signals
Spectral features extracted from 1-D RF signals	Temporal features extracted from 1-D RF signals	Several features combined extracted from 1-D RF signals

Machine learning models

1-NN DTW	MLP	FCN
ResNet	ROCKET	MINIROCKET
MultiRocket (with ROCKET kernels)	MultiRocket (with MINIROCKET kernels)	Transformer
CatBoost	XGBoost	LightGBM
SVM	Logistic Regression	

Figure 3.9.: Diagram showing all computed input signal combinations for muscle contraction state classification signals (adapted from [2]).

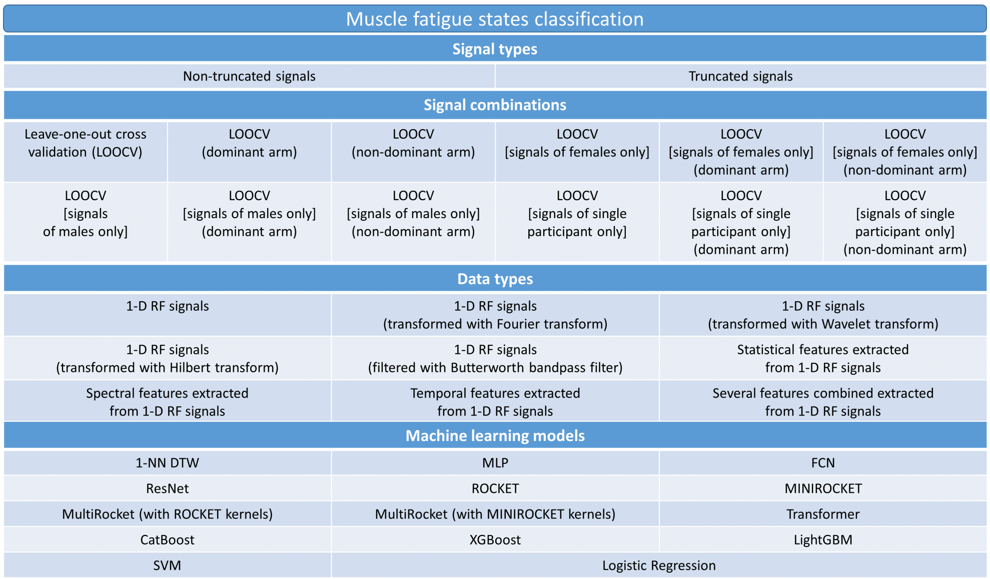
3.3.5.2. Muscle fatigue states evaluation schemes

For muscle fatigue state classifications, this work considered twelve different evaluation modes to compare the impact of a variety of signals and their properties on the results. These were the following training modes:

1. Study 1: LOOCV on all signals
(100.00 % of all signals)
2. Study 1: LOOCV on signals from the dominant arm only
(52.48 % of all signals)

3. Study 1: LOOCV on signals from the non-dominant arm only (47.52 % of all signals)
4. Study 1: LOOCV on signals from female subjects only (33.44 % of all signals)
5. Study 1: LOOCV on signals from male subjects only (66.56 % of all signals)
6. Study 1: LOOCV on signals from the dominant arm of female subjects only (17.54 % of all subjects)
7. Study 1: LOOCV on signals from the non-dominant arm of female subjects only (17.67 % of all subjects)
8. Study 1: LOOCV on signals from the dominant arm of male subjects only (34.94 % of all subjects)
9. Study 1: LOOCV on signals from the non-dominant arm of male subjects only (31.62 % of all subjects)
10. Study 2: LOOCV on signals from a single subject only (100.00 % of all signals)
11. Study 2: LOOCV on signals from the dominant arm of a single subject only (54.89 % of all signals)
12. Study 2: LOOCV on signals from the non-dominant arm of a single subject only (45.11 % of all signals)

Figure 3.10 shows a hierarchical diagram illustrating all computed input signal combinations for the muscle fatigue state classifications.



Muscle fatigue states classification					
Signal types					
Non-truncated signals			Truncated signals		
Signal combinations					
Leave-one-out cross validation (LOOCV)	LOOCV (dominant arm)	LOOCV (non-dominant arm)	LOOCV [signals of females only]	LOOCV [signals of females only] (dominant arm)	LOOCV [signals of females only] (non-dominant arm)
LOOCV [signals of males only]	LOOCV [signals of males only] (dominant arm)	LOOCV [signals of males only] (non-dominant arm)	LOOCV [signals of single participant only]	LOOCV [signals of single participant only] (dominant arm)	LOOCV [signals of single participant only] (non-dominant arm)
Data types					
1-D RF signals		1-D RF signals (transformed with Fourier transform)		1-D RF signals (transformed with Wavelet transform)	
1-D RF signals (transformed with Hilbert transform)		1-D RF signals (filtered with Butterworth bandpass filter)		Statistical features extracted from 1-D RF signals	
Spectral features extracted from 1-D RF signals		Temporal features extracted from 1-D RF signals		Several features combined extracted from 1-D RF signals	
Machine learning models					
1-NN DTW		MLP		FCN	
ResNet		ROCKET		MINIROCKET	
MultiRocket (with ROCKET kernels)		MultiRocket (with MINIROCKET kernels)		Transformer	
CatBoost		XGBoost		LightGBM	
SVM		Logistic Regression			

Figure 3.10.: Diagram illustrating all computed input signal combinations for muscle fatigue state classifications (adapted from [2]).

Note that not all permutational combinations of the diagrams depicted in Figure 3.9 and Figure 3.10 were possible in practice. For instance, the 1-NN DTW algorithm and models of the ROCKET family were not applicable for extracted features.

3.4. Results

3.4.1. Dimensionality reduction for signals of muscle contraction experiments

This section provides 2-D PCA and t-SNE plots colored by various properties for muscle contraction signals stemming from the same transducer position. In all plots, each dot represents a single Amplitude scan (A-Scan).

3.4.1.1. Principal component analysis

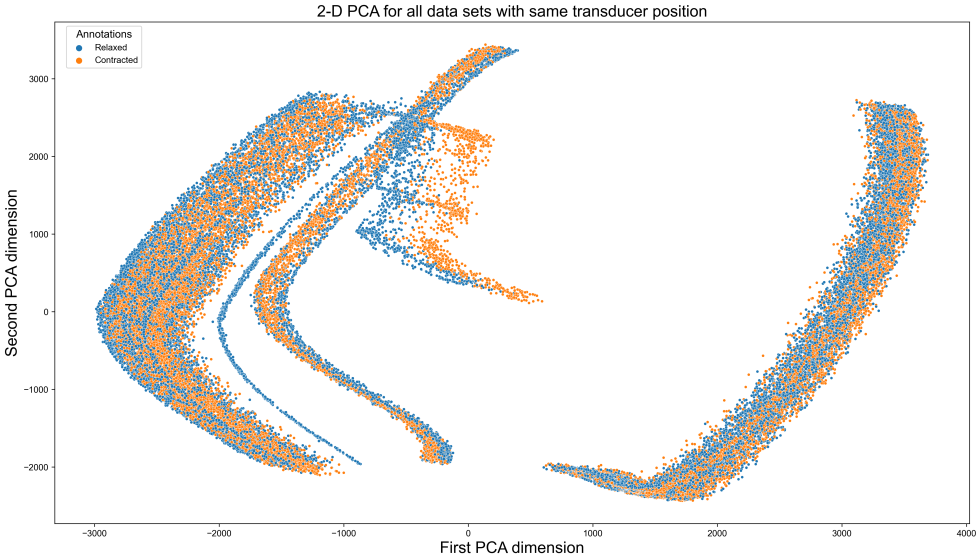


Figure 3.11.: 2-D PCA of all muscle contraction signals stemming from the same transducer position colored by annotation.

Figure 3.11 and Figure 3.12 show 2-D PCA visualizations (see Section 2.7.3.1) of the signal distribution of all A-scans stemming from datasets with the same transducer position. Figure 3.11 was colored according to annotations, while Figure 3.12 was colored according to datasets. In both visualizations the signals tend to group much stronger w.r.t. the datasets they belong to instead of the categories they have been annotated with.

3.4.1.2. t-distributed stochastic neighbor embedding

Figure 3.13 and Figure 3.14 show t-SNE visualizations (see Section 2.7.3.2) illustrating the low-dimensional signal distribution of all A-scans stemming from datasets with the same transducer position. Figure 3.13 was colored according to annotations, while Figure 3.14 was colored according to datasets. In contrast to the PCA plots depicted above, in these figures the signals tend to group much stronger w.r.t. the categories they have been annotated with instead of the datasets they belong to.

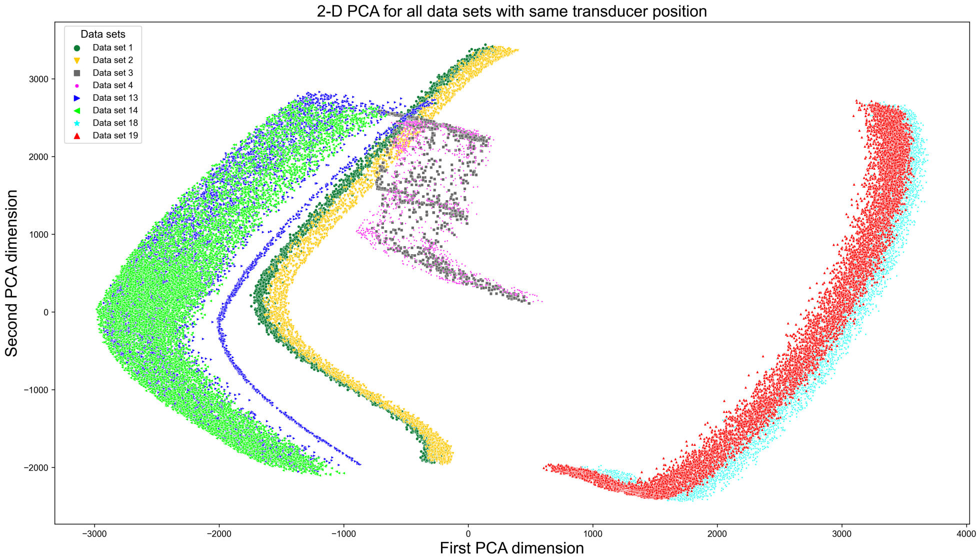


Figure 3.12.: 2-D PCA of all muscle contraction signals stemming from the same transducer position colored by dataset.

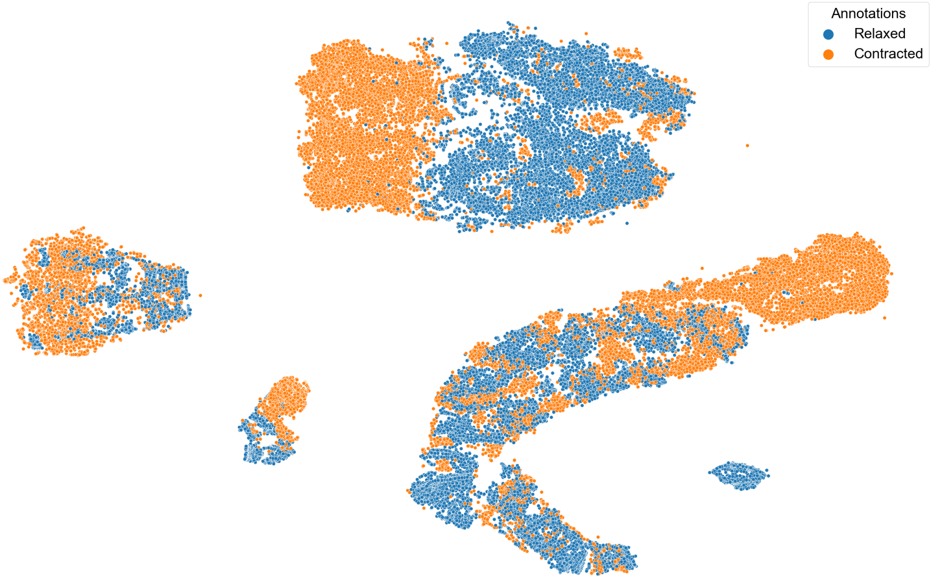


Figure 3.13.: 2-D t-SNE of all muscle contraction signals stemming from the same transducer position colored by annotation (adapted from [2]).

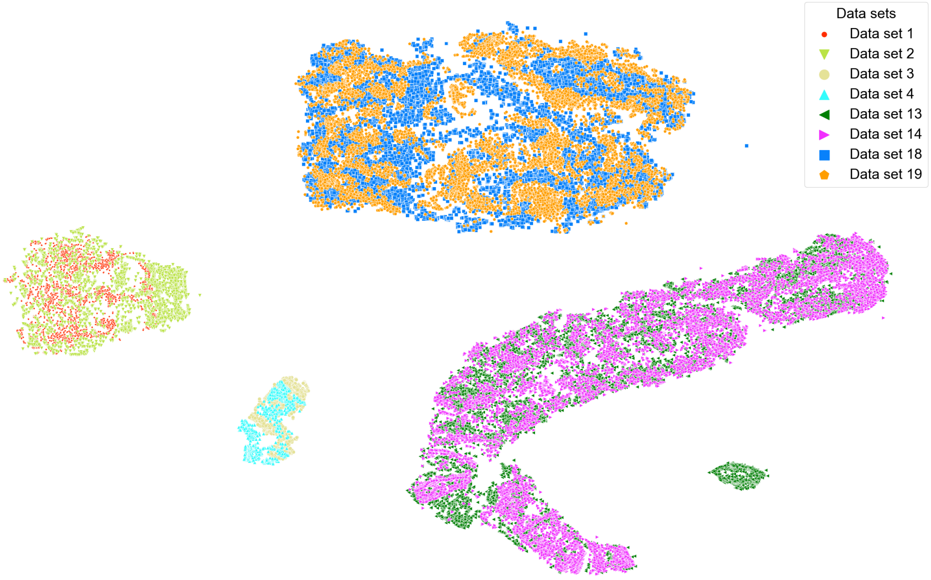


Figure 3.14.: 2-D t-SNE of all muscle contraction signals stemming from the same transducer position colored by dataset (adapted from [2]).

3.4.2. Achieved accuracies and training speed for muscle contraction state classifications

Table 3.5 summarizes the five best performing data type and ML model combinations based on the achieved average F_1-scores, while Figure 3.15 shows the rounded training evaluation times (in hours) for models trained on non-truncated raw or transformed muscle contraction signals or extracted features. For better interpretability, the results are illustrated with a logarithmic scale.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Hilbert transformed A-scans	88	0.17	no
MLP	Hilbert transformed A-scans	88	6.66	no
SVM	Fourier transformed A-scans	87	0.09	no
MLP	Fourier transformed A-scans	87	6.12	no
SVM	Wavelet transformed A-scans	86	0.12	no

Table 3.5.: Five best performing model / data type combinations for a validation on all muscle contraction signals.

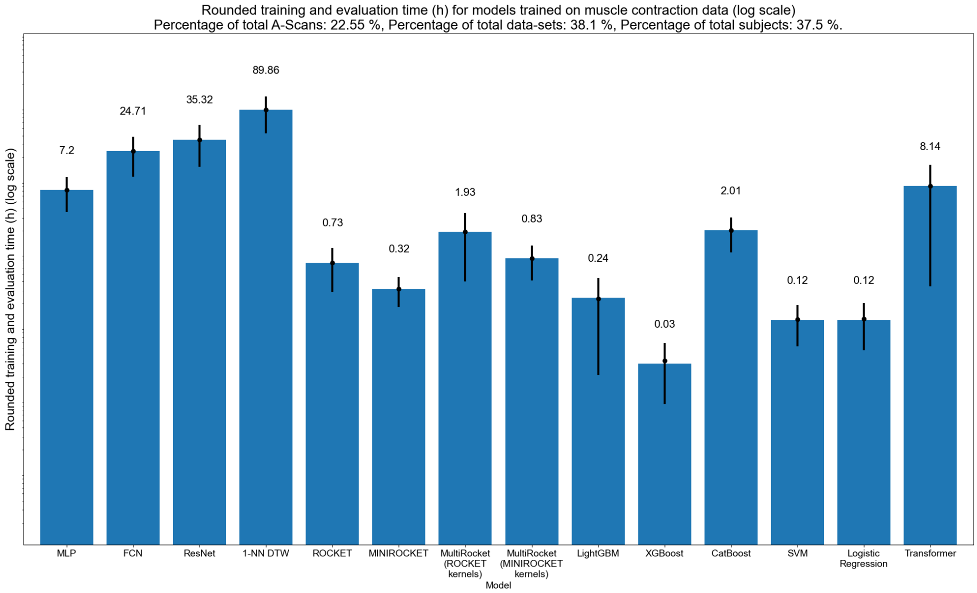


Figure 3.15.: Training and evaluation times for all data type / ML model combinations of muscle contraction signals stemming from the same person and the same ultrasound transducer position.

3.4.3. Dimensionality reduction for signals of muscle fatigue experiments

This section provides 2-D PCA and t-SNE plots colored according to various properties for muscle fatigue signals stemming from study 1 and study 2. In all plots, each dot represents a single A-Scan.

3.4.3.1. Principal component analysis

Study 1

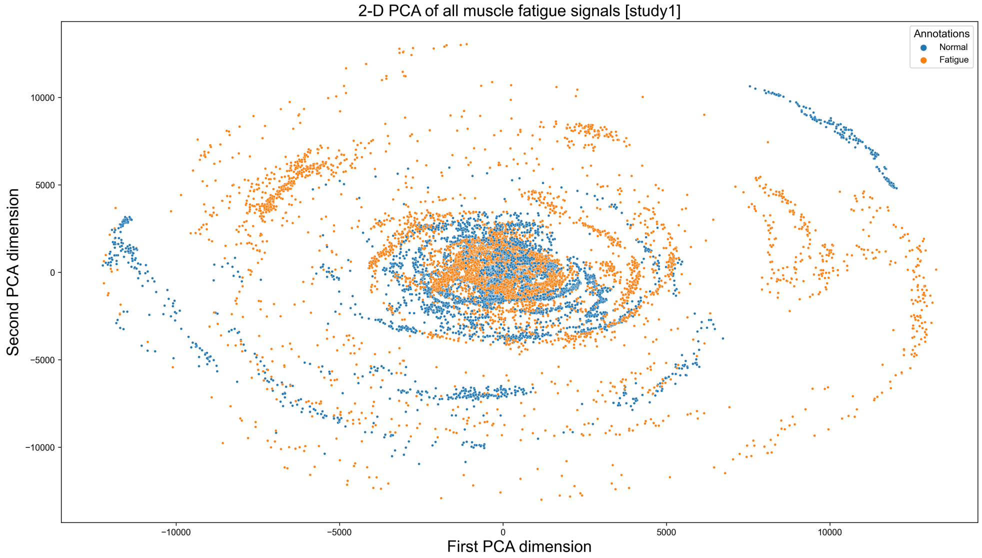


Figure 3.16.: 2-D PCA of all muscle fatigue signals stemming from study 1 colored by annotation.

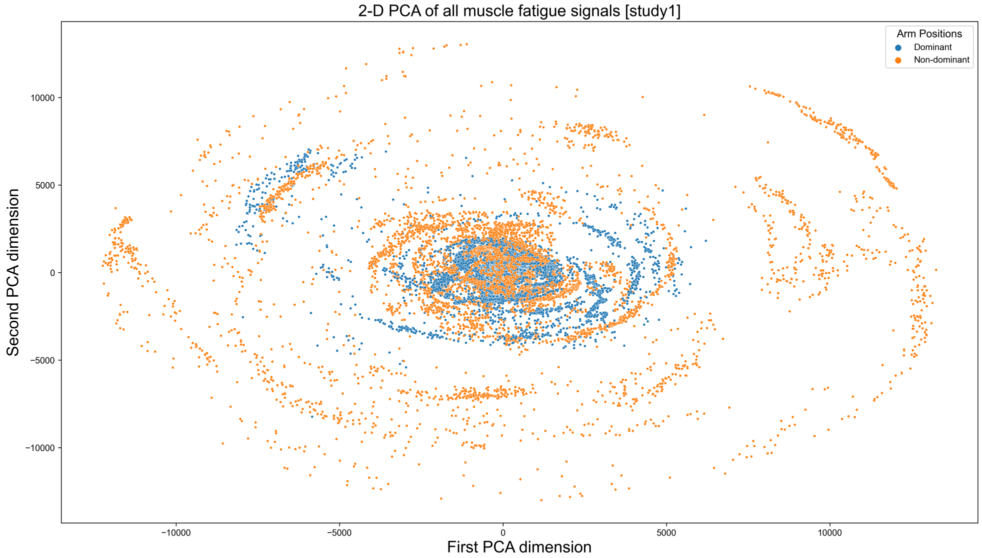


Figure 3.17.: 2-D PCA of all muscle fatigue signals stemming from study 1 colored by arm position (dominant vs. non-dominant arm).

Figure 3.16, Figure 3.17, Figure 3.18, Figure 3.19 and Figure 3.20 show 2-D PCA plots of signals stemming from study 1 colored according to annotations, arm positions, sex, maximum lifted weight and subject ID, respectively. These figures illustrate vividly that the signals tend to group together most strongly

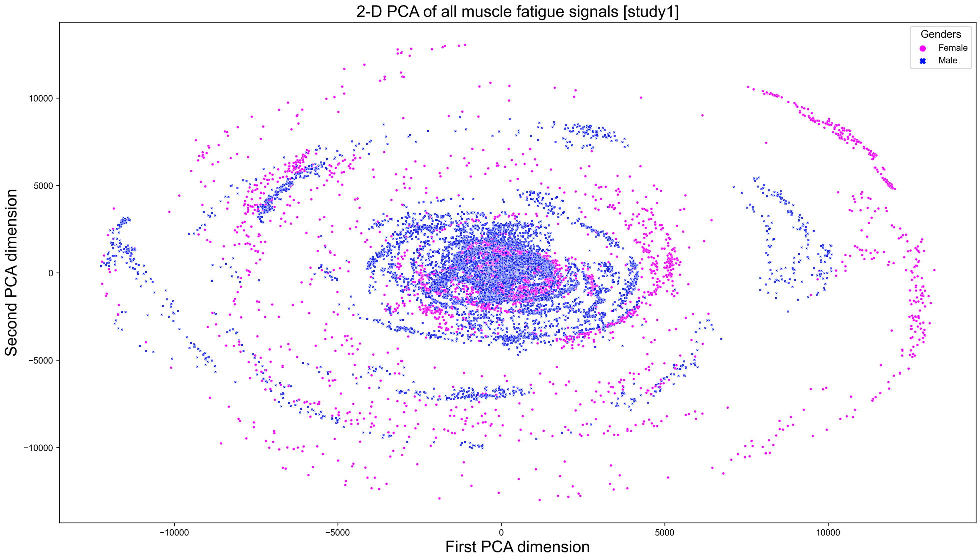


Figure 3.18.: 2-D PCA of all muscle fatigue signals stemming from study 1 colored by sex (male vs. female).

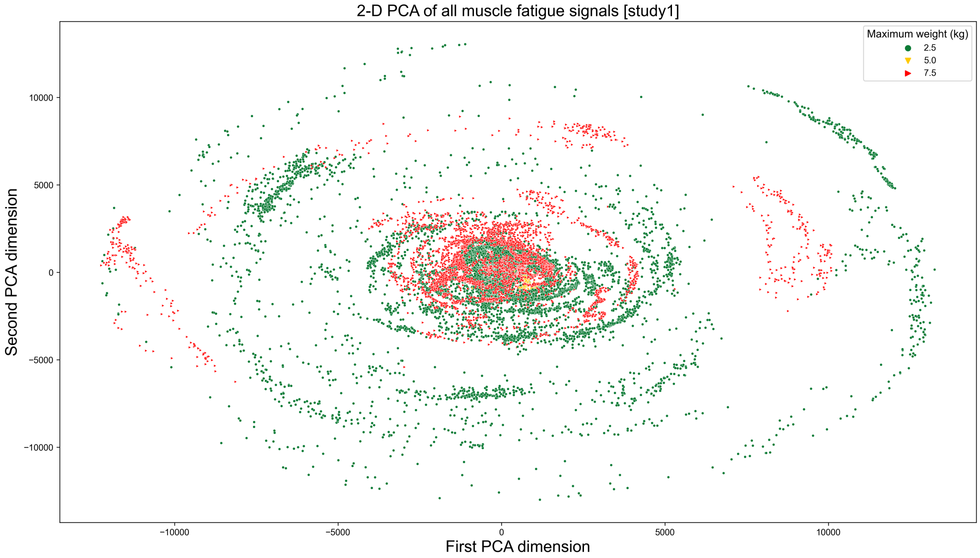


Figure 3.19.: 2-D PCA of all muscle fatigue signals stemming from study 1 colored by maximally lifted weight (2.5 kg vs. 5.0 kg vs. 7.5 kg).

according to arm position (i.e. dominant arm vs. non-dominant arm). Signal groupings according to other properties are much less pronounced but still display strong tendencies.

Study 2

Figure 3.21 and Figure 3.22 show 2-D PCA plots of signals stemming from study 2 colored according to annotations and arm positions, respectively.

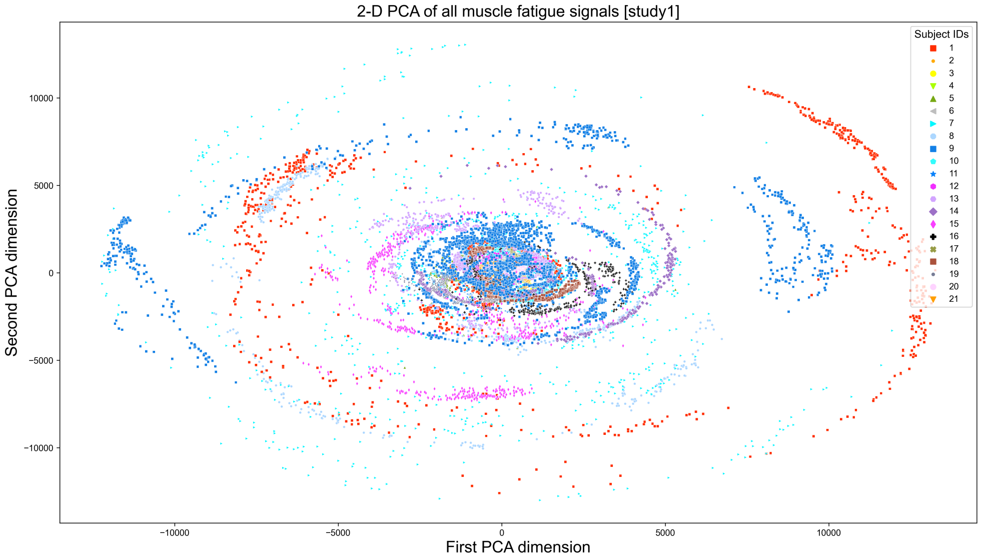


Figure 3.20.: 2-D PCA of all muscle fatigue signals stemming from study 1 colored by subject.

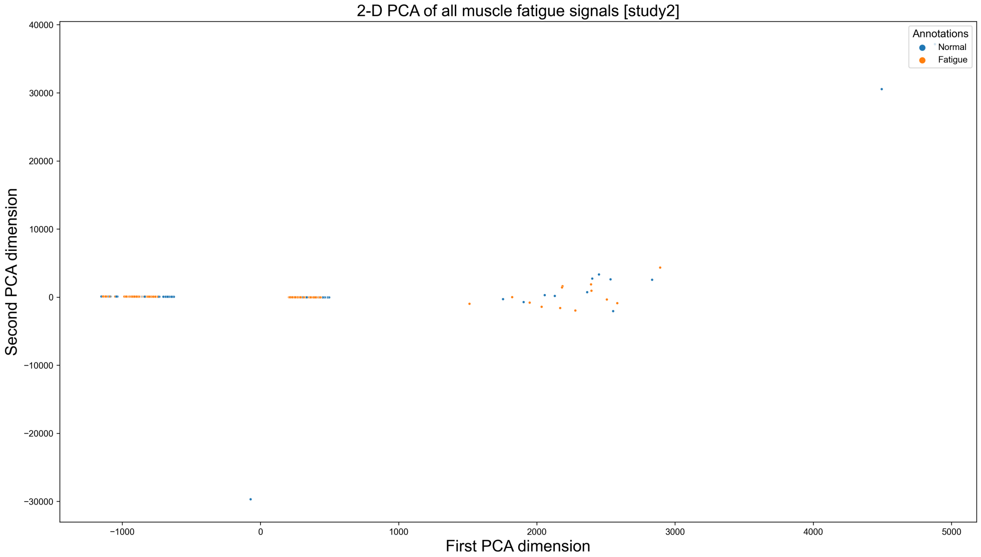


Figure 3.21.: 2-D PCA of all muscle fatigue signals stemming from study 2 colored by annotation.

These figures show a very distinct distribution pattern, which does not allow any conclusions concerning strict signal groupings. Both figures show string grouping tendencies but no figure display a clear separation of different categories.

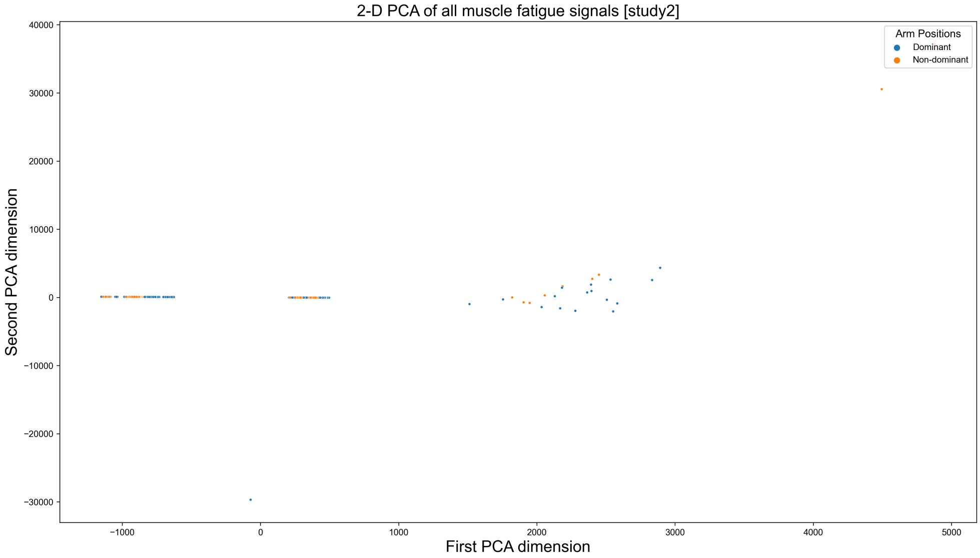


Figure 3.22.: 2-D PCA of all muscle fatigue signals stemming from study 2 colored by arm position (dominant vs. non-dominant arm).

3.4.3.2. t-distributed stochastic neighbor embedding

Study 1

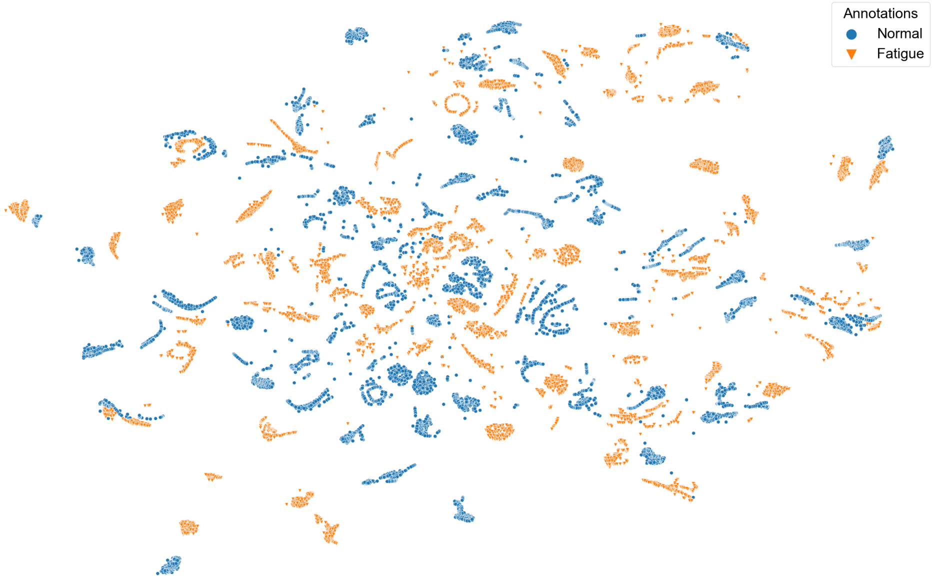


Figure 3.23.: 2-D t-SNE of all muscle fatigue signals stemming from study 1 colored by annotation (adapted from [2]).

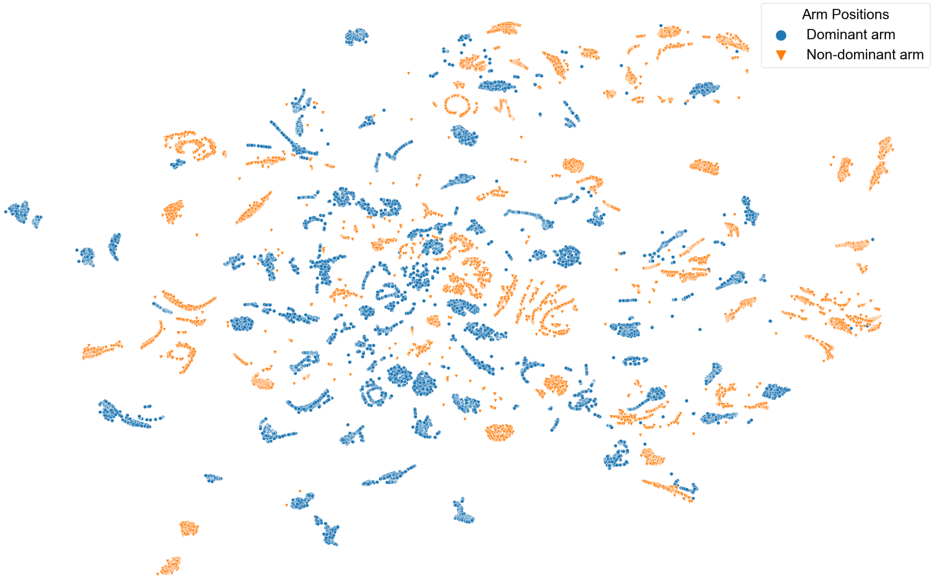


Figure 3.24.: 2-D t-SNE of all muscle fatigue signals stemming from study 1 colored by arm position (dominant vs. non-dominant arm) (adapted from [2]).

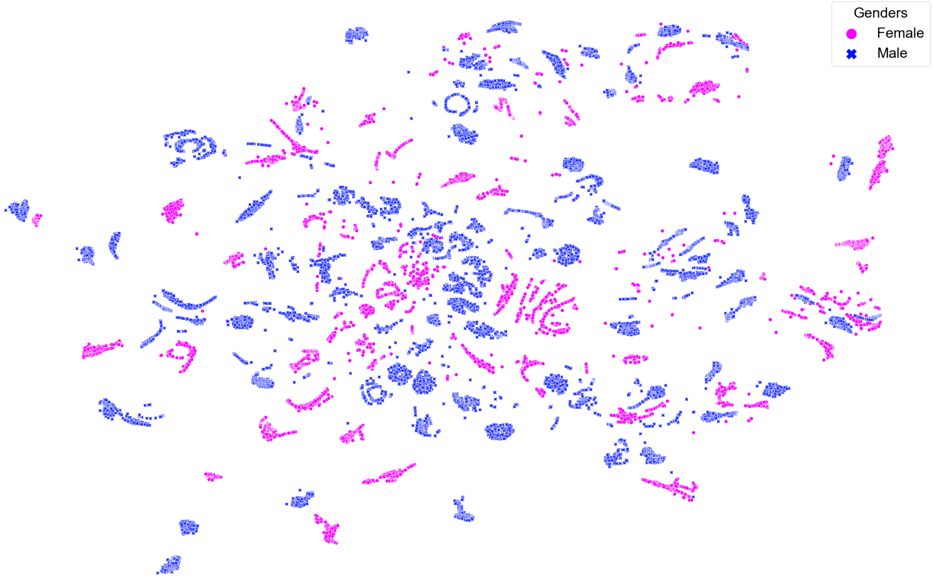


Figure 3.25.: 2-D t-SNE of all muscle fatigue signals stemming from study 1 colored by sex (male vs. female) (adapted from [2]).

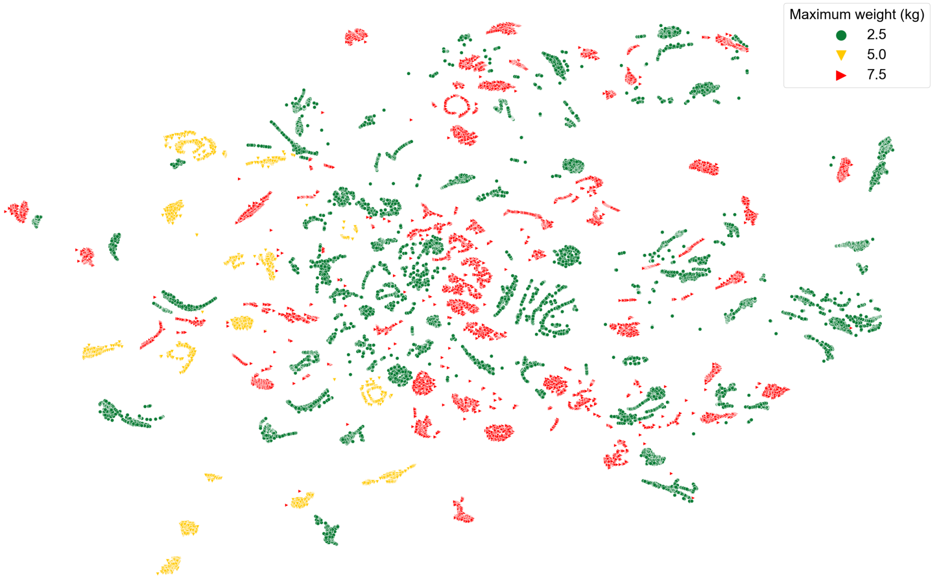


Figure 3.26.: 2-D t-SNE of all muscle fatigue signals stemming from study 1 colored by maximally lifted weight (2.5 kg vs. 5.0 kg vs. 7.5 kg) (adapted from [2]).

Figure 3.23, Figure 3.24, Figure 3.25, Figure 3.26 and Figure 3.27 show 2-D t-SNE visualizations of the muscle fatigue state signal distribution from study 1. Each dot represents a single A-Scan. Figure 3.23 is color-coded according to the muscle state (normal vs. fatigue) and Figure 3.24 is color-coded according to the arm the signals stem from (dominant vs. non-dominant). Figure 3.25 is

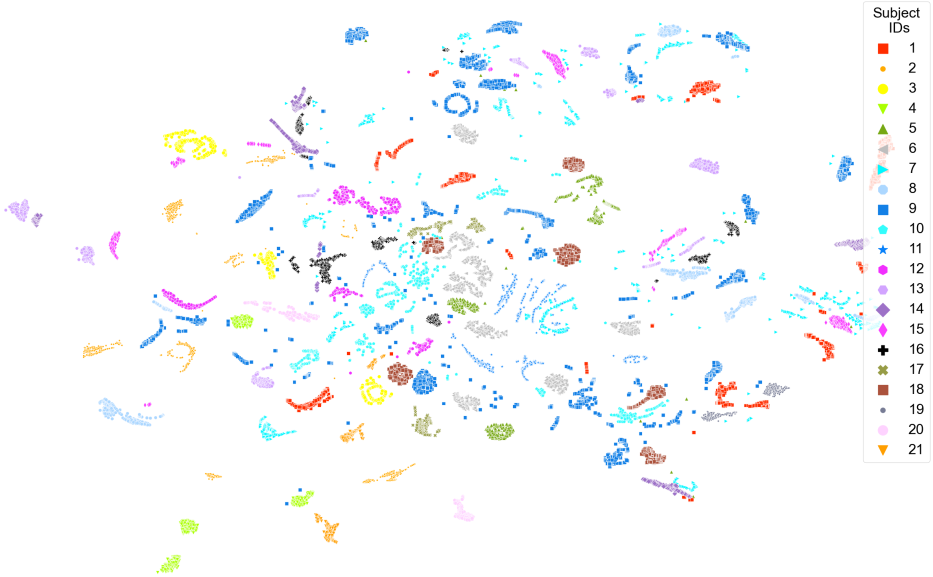


Figure 3.27.: 2-D t-SNE of all muscle fatigue signals stemming from study 1 colored by subject (adapted from [2]).

color-coded according to sex (female vs. male), while Figure 3.26 is color-coded according to the maximum weight lifted (2.5 kg, 5.0 kg, or 7.5 kg). Figure 3.27 is color-coded by subjects. Figure 3.23 shows that there is no strict grouping of the signals according to muscle states, even though a slight tendency is visible. Figure 3.24 illustrates that the signals tend to group according to arm positions. However, this grouping is not very strict and shows only slight tendencies instead of rigorous borders. Figure 3.25 shows that the signals tend to group according to the sex of each subject. However, this grouping is also not very strict and shows only slight tendencies instead of rigorous borders. Figure 3.26 shows a slight tendency of the signals to group according to the maximum weight they have been annotated with. Figure 3.27 shows that the signals have a very strong tendency to group together according to the subject they belong to.

Study 2

Figure 3.28 and Figure 3.29 show 2-D t-SNE visualizations of the signal distribution from study 2 of the muscle fatigue states classification. Figure 3.28 is color-coded according to the muscle state (normal vs. fatigue) and Figure 3.29 is color-coded according to the arm position (dominant vs. non-dominant). Figure 3.28 shows that the signals of study 2 only have a slight tendency to group according to the muscle state they belong to, while Figure 3.29 shows that the signals of study 2 have a strong tendency to group according to the arm position they have been annotated with.
All t-SNE plots shown above for study 1 and study 2 supported the

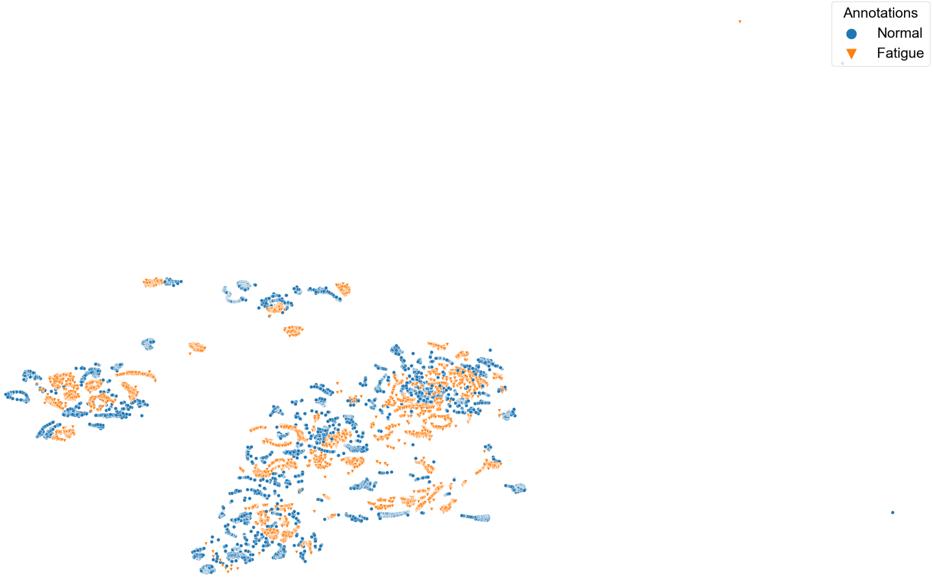


Figure 3.28.: 2-D t-SNE of all muscle fatigue signals stemming from study 2 colored by annotation (adapted from [2]).

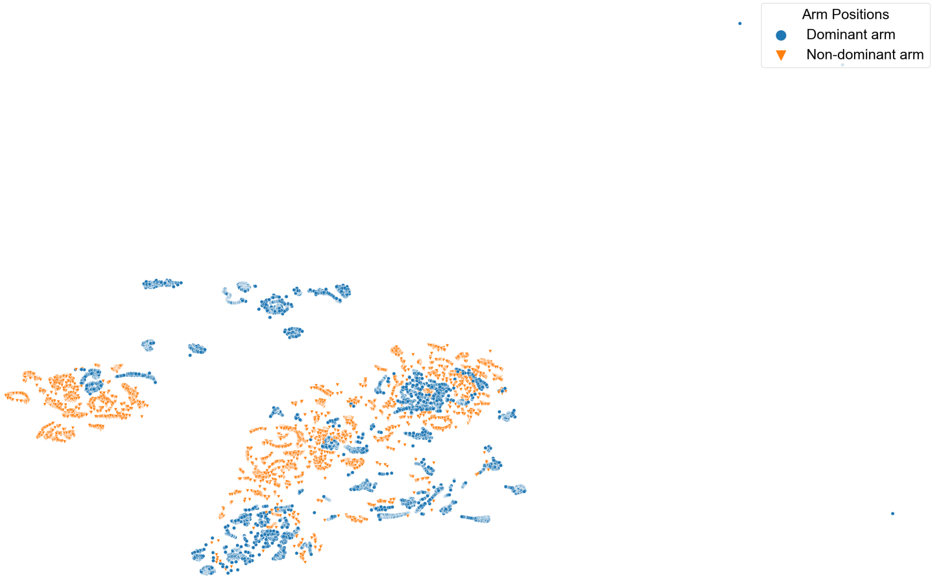


Figure 3.29.: 2-D t-SNE of all muscle fatigue signals stemming from study 2 colored by arm position (dominant vs. non-dominant arm) (adapted from [2]).

construction of several evaluation schemes as described in Section 3.3.5.2.

3.4.4. Achieved accuracies and training speed for muscle fatigue state classifications

This section presents the five best performing model / data type combinations and the corresponding times needed for training and evaluation for each scheme presented in Section 3.3.5.2. Note that the training and evaluation times for models using extracted features do *not* include the time needed to extract features as this was a separate process of the ML pipeline. Appendix A.1 provides a complete and detailed overview of the achieved F_1-scores stemming from all data type and ML model combinations. Appendix A.2 provides a complete and detailed overview of the training and evaluation times needed for each data type / model combination.

3.4.4.1. Leave-one-out cross validation

Table 3.6 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals.

Model	**Data type**	**Average F_1-score (%)**	**Time for training and evaluation (h)**	**Signals truncated**
SVM	Wavelet transformed A-scans	82	5.57	no
SVM	Raw A-scans	77	7.79	no
ROCKET	Wavelet transformed A-scans	77	10.50	no
MultiRocket (with ROCKET kernels)	Wavelet transformed A-scans	77	11.94	no
SVM	All features combined	77	0.55	no

Table 3.6.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals.

3.4.4.2. Leave-one-out cross validation (dominant arm)

Table 3.7 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the dominant arm.

Model	**Data type**	**Average F_1-score (%)**	**Time for training and evaluation (h)**	**Signals truncated**
SVM	Wavelet transformed A-scans	84	0.33	no
SVM	Raw A-scans	83	0.50	no
Logistic Regression	Raw A-scans	82	1.87	no
SVM	All features combined	80	0.07	no
Transformer	Wavelet transformed A-scans	80	20.93	no

Table 3.7.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the dominant arm.

3.4.4.3. Leave-one-out cross validation (non-dominant arm)

Table 3.8 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the non-dominant arm.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Wavelet transformed A-scans	77	0.08	yes
SVM	Spectral features	76	0.08	yes
MultiRocket (with ROCKET kernels)	Wavelet transformed A-scans	76	2.35	no
ROCKET	Wavelet transformed A-scans	76	1.47	no
SVM	Wavelet transformed A-scans	72	0.24	no

Table 3.8.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the non-dominant arm.

3.4.4.4. Leave-one-out cross validation (female)

Table 3.9 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from female subjects only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
Logistic Regression	All features combined	77	0.05	no
Logistic Regression	Spectral features	76	0.04	no
ROCKET	Wavelet transformed A-scans	76	0.49	no
ROCKET	Fourier transformed A-scans	76	0.49	no
Transformer	Spectral features	76	6.16	no

Table 3.9.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from female subjects only.

3.4.4.5. Leave-one-out cross validation (female and dominant arm)

Table 3.10 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the dominant arm of female subjects only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
Logistic Regression	Spectral features	86	0.01	no
Logistic Regression	All features combined	84	0.01	no
Transformer	Spectral features	81	1.79	no
Logistic Regression	Temporal features	76	0.00	no
Logistic Regression	Fourier transformed A-scans	76	0.02	no

Table 3.10.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the dominant arm of female subjects only.

3.4.4.6. Leave-one-out cross validation (female and non-dominant arm)

Table 3.11 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the non-dominant arm of female subjects only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Wavelet transformed A-scans	75	0.01	no
Logistic Regression	Fourier transformed A-scans	73	0.02	no
ROCKET	Fourier transformed A-scans	73	0.10	no
Transformer	Wavelet transformed A-scans	73	2.99	no
Logistic Regression	Spectral features	71	0.01	no

Table 3.11.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the non-dominant arm of female subjects only.

3.4.4.7. Leave-one-out cross validation (male)

Table 3.12 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from male subjects only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Wavelet transformed A-scans	84	0.84	no
Transformer	Wavelet transformed A-scans	80	55.26	no
SVM	Raw A-scans	79	1.65	no
Logistic Regression	Spectral features	79	0.24	no
Logistic Regression	All features combined	79	0.28	no

Table 3.12.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from male subjects only.

3.4.4.8. Leave-one-out cross validation (male and dominant arm)

Table 3.13 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the dominant arm of male subjects only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Wavelet transformed A-scans	86	0.01	no
SVM	Raw A-scans	84	0.01	no
Transformer	Wavelet transformed A-scans	83	1.79	no
Logistic Regression	Raw A-scans	82	0.00	no
Logistic Regression	Wavelet transformed A-scans	82	0.02	no

Table 3.13.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the dominant arm of male subjects only.

3.4.4.9. Leave-one-out cross validation (male and non-dominant arm)

Table 3.14 shows the five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the non-dominant arm of male subjects only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	All features combined	79	0.02	yes
SVM	Spectral features	78	0.02	yes
SVM	All features combined	78	0.01	no
Transformer	Temporal features	78	12.03	no
SVM	Wavelet transformed A-scans	77	0.06	no

Table 3.14.: Five best performing model / data type combinations for a leave-one-out cross validation on all muscle fatigue signals from the non-dominant arm of male subjects only.

3.4.4.10. Leave-one-out cross validation of single subject

Table 3.15 shows the five best performing model / data type combinations for a leave-one-out cross validation on muscle fatigue signals from a single subject only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
Logistic Regression	Statistical features	70	0.02	no
SVM	All features combined	70	0.31	no
Transformer	Temporal features	67	27.26	no
SVM	Fourier transformed A-scans	66	0.90	no
1-NN DTW	Raw A-scans	66	36.04	no

Table 3.15.: Five best performing model / data type combinations for a leave-one-out cross validation on muscle fatigue signals from a single subject only.

3.4.4.11. Leave-one-out cross validation of single subject (dominant arm)

Table 3.16 shows the five best performing model / data type combinations for a leave-one-out cross validation on muscle fatigue signals of the dominant arm from a single subject only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Wavelet transformed A-scans	78	0.11	no
Logistic Regression	Statistical features	77	0.01	no
SVM	All features combined	75	0.04	no
SVM	Fourier transformed A-scans	74	0.12	no
SVM	Raw A-scans	70	0.20	no

Table 3.16.: Five best performing model / data type combinations for a leave-one-out cross validation on muscle fatigue signals of the dominant arm from a single subject only.

3.4.4.12. Leave-one-out cross validation of single subject (non-dominant arm)

Table 3.17 shows the five best performing model / data type combinations for a leave-one-out cross validation on muscle fatigue signals of the non-dominant arm from a single subject only.

Model	Data type	Average F_1-score (%)	Time for training and evaluation (h)	Signals truncated
SVM	Temporal features	72	0.01	no
Transformer	Temporal features	72	4.55	no
SVM	All features combined	70	0.02	no
Logistic Regression	Temporal features	70	0.00	no
1-NN DTW	Bandpass filtered A-scans	70	8.65	no

Table 3.17.: Five best performing model / data type combinations for a leave-one-out cross validation on muscle fatigue signals of the non-dominant arm from a single subject only.

3.5. Discussion of muscle state classifications

3.5.1. Muscle contractions

3.5.1.1. Dimensionality reduction techniques

In the 2-D PCA visualizations shown in Section 3.4.1.1, the signals tend to group much stronger w.r.t. the datasets they belong to instead of the categories they have been annotated with. In contrast to the PCA plots, the 2-D t-SNE visualizations presented in Section 3.4.1.2 show that the signals tend to group much stronger w.r.t. the categories they have been annotated with instead of the datasets they belong to. The t-SNE method is suitable for non-linear signal distributions, while the PCA method is not [136]. This explains discrepancies between the 2-D visualizations of the former and the latter. The t-SNE visualization served as foundation for the main hypothesis that a robust classification of the acquired signals based on the respective annotations is possible.

3.5.1.2. Machine Learning

A SVM (see Section 2.7.4.5) model based on Hilbert transformed A-scans achieved an average F_1-score of ca. 88 % in less than 10 minutes for training and evaluation. This SVM model exceeded the performance of more recent ANN models and even the classic 1-NN DTW algorithm, which has been the de facto TSC benchmark for decades [61]. All other deployed ML models yielded worse performances w.r.t. achieved F_1-scores or training and evaluation times.

3.5.1.3. Conclusion

A remarkable result is that a SVM (see Section 2.7.4.5) model based on Hilbert transformed A-scans even outperformed most recent ANNs in terms of speed and accuracy. This result shows that traditional algorithms can still be deployed to achieve superior results with a very low ecological footprint and also paves the way for real-life applications allowing wearable devices to classify different muscle contraction states based on 1-D SMG signals in a matter of minutes. To this end, the popular software library *LibSVM*, which serves as foundation for many implementations aimed for mobile systems, could be a core component of implementations for *Android* or *iOS* in future works [193].

3.5.2. Muscle fatigue

3.5.2.1. Dimensionality reduction techniques

Section 3.4.3.1 presents 2-D PCA visualizations for study 1 and study 2 of the muscle fatigue states classifications. The signal distribution plots of study 1 illustrate that the signals tend to group together most strongly according to arm position (i.e. dominant arm vs. non-dominant arm). Signal groupings according to other properties are much less pronounced but still display strong tendencies. The signal distribution plots of study 2 show a very distinct distribution pattern, which does not allow any conclusions concerning strict signal groupings. Note that PCA is not considered a suitable method for non-linear signal distributions [136]. Hence, Section 3.4.3.2 shows plots created with the t-SNE algorithm, which has been created for both linear and non-linear signal distributions [136]. The 2-D t-SNE plots of signals from study 1 show that they have a very strong tendency to group together according to their associated subject, while the t-SNE plots of study 2 show that the signals have a strong tendency to group according to their associated arm position. Even though the t-SNE plots do not show a strict grouping of the signals according to different muscle states, the displayed grouping tendencies served as foundation for the main hypothesis that robust classifications of different muscle states are possible.

3.5.2.2. Machine Learning

Regardless of the evaluation scheme (see Section 3.3.5.2) and the data type / ML model combination, muscle fatigue state classifications based on signals stemming from the *dominant* arm **always** achieved superior results in comparison to classifications based on signals stemming from both arms or from the non-dominant arm. A likely reason for this is the fact that muscles from the dominant arm are usually more pronounced, which most probably resulted in acquired signals less affected by inhomogeneities. US B-Mode imaging “has been widely used [...] to evaluate the morphological and mechanical properties of [muscles] and [tendons, and] a regular assessment of such properties has great potential, namely for testing the response to training, detecting athletes at higher risks of injury, screening athletes for structural abnormalities related to current or future musculoskeletal complaints, and monitoring their return to sport after a musculoskeletal injury” [194]. Hence, it is plausible to assume that significant amplitude differences between signals stemming from dominant and non-dominant arms played an important role for the achieved classification results.

An observation holding true for all evaluation schemes is that the ML models SVM (see Section 2.7.4.5) and LR (see Section 2.7.4.1) **always** outperformed all other ML models consistently. These models also required much less time for training and evaluation than many other approaches (see Appendix A.2). This was a remarkable result and proved that highly sophisticated and comparatively recent ANNs or GBMs are not necessarily faster or better than straightforward approaches that have been in use for decades.

Wavelet transformed signals (see Section 2.3.2.2), a combination of all extracted features, statistical features (see Section 2.3.3.1, temporal features (see Section 2.3.3.2) and spectral features (see Section 2.3.3.3) are the best performing data types for 6, 3, 1, 1 and 1 out of 12 compared evaluation schemes, respectively. Thus, Wavelet transformed signals serving as input data for a SVM model can be considered a reasonable benchmark combination for future 1-D US signal classifications.

Overall, the evaluation schemes based on signals stemming from the dominant arm and the same sex perform best. A LR model using extracted spectral features as input data and a SVM model using Wavelet transformed signals as input data both yielded an average F_1-score of 86 %, while finishing training and evaluation in less than 6 minutes. For the LOOCV evaluation scheme, a SVM model using Wavelet transformed signals from the dominant arm as input data yielded an average F_1-Score of 84 %, while finishing training and evaluation in only slightly more than 30 minutes. The performance of all data type / ML model combinations was slightly worse for ML models based on signals stemming from the second study, which contained only signals from a single male subject. Signals from a single subject are comparatively homogeneous, while signals from a larger variety of different subjects introduce more inhomogeneity. It is plausible to assume that these phenomena lead to better general classification results for the first study as ML models were more likely to deduce significant separations if presented with differently pronounced data. For the single subject scenario, a SVM model using Wavelet transformed signals from the dominant arm as input data performed best with an average F_1-Score of 78 %.

In most cases ML models based on raw 1-D A-scans outperformed ML models based on truncated A-scans. In the rare cases where the latter performed better, the difference to the second best performing model in terms of average F_1-score was never more than one percent point.

3.5.2.3. Conclusion

In general, fast and accurate machine-assisted classifications of different muscle fatigue states based on 1-D US signals are possible. The results indicate that grouping the acquired signals into categories separated by sex and arm position yields the best performing ML models. The performance of ML models based on signals from a variety of different subjects is slightly better than the performance of ML models based on signals from a single male subject only. Table 3.18 presents a summary of the results of all evaluation schemes for muscle fatigue states classifications.

Even though these results are very promising, the presented data type / ML model combinations are not yet suitable for clinical applications. They are, however, accurate enough to be integrated in devices meant for recreational

Evaluation mode	**ML model**	**Data type**	F_1**-score**	**Time for evaluation and training (in minutes)**
LOOCV	SVM	Wavelet transformed A-scans	82	334
LOOCV (dominant arm)	SVM	Wavelet transformed A-scans	84	20
LOOCV (non-dominant arm)	SVM	Combination of all possible features	77	<5
LOOCV (female)	Logistic Regression	Combination of all possible features	77	<5
LOOCV (female) [dominant arm]	Logistic Regression	Spectral features	86	<5
LOOCV (female) [non-dominant arm]	Logistic Regression	Wavelet transformed A-Scans	75	<5
LOOCV (male)	SVM	Wavelet transformed A-scans	84	50
LOOCV (male) [dominant arm]	SVM	Wavelet transformed A-scans	86	5
LOOCV (male) [non-dominant arm]	SVM	Combination of all possible features (of truncated signals)	79	<5
LOOCV (single subject 09)	Logistic Regression	Statistical features	70	<5
LOOCV (single subject 09) [dominant arm]	SVM	Wavelet transformed A-scans	78	7
LOOCV (single subject 09) [non-dominant arm]	SVM	Temporal features	72	<5

Table 3.18.: Summary of the results of all evaluation schemes for muscle fatigue states classifications.

purposes, such as wearable solutions tracking the progress of fitness programs.

Chapter 4

Detection of epiphyseal radius bone closure

The distal growth plate fusion in the ulna and radius bones is commonly examined via X-ray imaging to determine bone ages. However, X-ray imaging has the disadvantage of being based on ionizing radiation, which necessitates reliable alternatives based on non-ionizing radiation, such as US. This chapter presents a low-cost, portable system relying on ML models based on 1-D US signals performing robust binary classifications to determine epiphyseal radius bone closures from signals of girls and women aged 9 to 24 years.

The developed system detects the presence or absence of epiphyseal radius bone closures by moving custom-designed US array transducers along the forearm, which measure reflection and transmission signals. Using ML approaches, the developed system can detect the distal growth plate fusion of the radius bone and the end of bone growth with a high accuracy.

Partial results of the presented work in this chapter have been published in [1].

Contents

4.1. Introduction

Bone age is a metric to determine skeletal maturity. It is defined by the age that corresponds to the level of bone maturity in an examined subject and can differ from chronological age, which is defined as the age between birth and present [195]. The maturation of skeletal bones "is based on the activation and interaction of a complex series of physiological mechanisms. This process is characterized by a predictable sequence of development and progression of ossification centers. Each bone segment begins its maturation first in the primary ossification center and then, through different stages of enlargement and remodeling, reaches the final shape". Certain bones, such as long bones, present with several centers of maturation called *epiphysis* [195]. Radius and ulna are long bones found in the forearm that stretches from the elbow to the smallest finger. The ulna is usually slightly longer than the radius, but the radius is thicker. Certain circumstances, such as unusual levels of growth hormones or Insulin-like growth factor-1, a deficit of thyroid hormones, an excess of corticosteroids and sex can strongly affect the process of skeletal maturation [195].

4.1.1. Applications of bone age determinations

Bone age determination is crucial to assess whether a subject presents with a stature that is deemed too short or too tall for a given chronological age or whether a subject suffers from impaired or accelerated growth, delayed or early puberty or the progression of several endocrine diseases.

In addition to the use cases presented above, hand and wrist radiographic images to determine bone ages are also used in nonmedical fields, such as sports or age assessments of asylum seekers. "However, bone age itself cannot be considered the only absolute and incontrovertible datum to define the chronological age [and] therefore, limits and accuracy of this examination in predicting chronological age, especially in relation to different ethnic groups and underlying diseases, need to be considered" [195]. The age determination in asylum seekers based on bone age assessments is especially controversial. The *European Academy of Paediatrics* even "strongly [recommends] all paediatricians in Europe not to participate in the process of age determinations in minor asylum seekers stating they are minors"; partly due to a lack of highly accurate bone age determination methods [196].

4.1.2. Bone age atlases

Numerous standardized methods have been developed in the past to evaluate skeletal maturity based on hand or wrist radiographs. The *Greulich-Pyle* method, the *Tanner-Whitehouse* method and the *Fels* method are the most representative [195]. All of these bone age atlases are based on radiographs acquired with X-ray imaging. "Between the years 1931 and 1942, the authors Greulich and Pyle evaluated hand and wrist radiographs acquired from about 1,000 white people from Cleveland (Ohio, USA)". The Tanner-Whitehouse method was developed based on data from 1,930 European children. This system has later been refined by moving from the *TW1* method to the subsequent methods *TW2* and *TW3*. The latter approach is based on data obtained from native North American children. The Fels method was developed by a study based on "a total of 13,823 serial [X-ray images] of the left hand and wrist. These images were [acquired] from 355 male and 322 female children born between the years 1928 and 1974, from the first month of life up to the age of 22 years". Although this approach is very accurate, it has been described as being "too complex" for daily use [195].

4.1.3. Alternative methods of bone age determination

The system *BoneXpert*, based on radiographs was introduced in 2008. Its algorithm has been "validated for different ethnic groups and for children with different endocrine disorders". However, this system does not take carpal bone evaluations into account and has been described as being opposed by local administrations due to higher costs compared to available methods [195]. Apart from methods based on radiographs, bone age assessment methods based on Magnetic resonance imaging (MRI) images of the left hand or the knee have also been proposed [197]. However, the acquisition of radiographs requires ionizing radiation and MRI is an expensive and less accessible technology. To address the shortcomings of these technologies, non-invasive *quantitative ultrasound (QUS)* approaches have been proposed in the past (see Section 4.2).

4.1.4. Challenges of bone age determination

Regardless of the underlying technology, such as radiographs, US or MRI, challenges of bone age determination remain. "Currently, hand and wrist X-ray is the gold standard to assess children's bone [ages]". However, they do have their shortcomings as "a proper assessment of bone age must always take into account differences between ethnic groups, sex, and any present pathological conditions" [195]. The Greulich and Pyle atlas, for example, has been described as being "imprecise and [it] should be used with caution when applied to Asian male and African female populations, particularly when aiming to determine chronological age for forensic [or] legal purposes" [198]. Another study states that "[all] current methods of assessing skeletal maturation are based primarily on a white population and are not necessarily generalizable to children of other ethnicities, particularly [from] African and certain Asian backgrounds. These limitations are even more important when bone ages are used for high-stake decisions. Further debate is needed on the risks and ethics associated with using bone age for nonmedical purposes. Many newer methods, which may be calibrated to specific populations, may perform better for a wider range of ethnicities, but more data are needed" [199].

4.2. Related works

First attempts of applying QUS methods for bone age determinations by performing B-Mode imaging of the femoral head articular cartilage (FHC) were proposed in 1995 [200]. However, the clinical use of approaches based on FHC signals was strongly discouraged in a 1998 publication, which reasoned that the sensitivity of these methods was "too low" [201]. A later study evaluated the QUS system BoneAge™ on 37 children and found that it was sufficiently accurate to determining their bone ages [26]. This system computes bone ages by combining US propagation speeds and the distance between emitter and receiver and comparing those values to a database containing sex and ethnicity reference data [202]. In two studies, the BoneAge™ system was evaluated on 152 and 65 subjects respectively and demonstrated a sufficient accuracy of bone age determinations [202, 203]. Regardless of these promising results, another publication found that the potential of US for the assessment of bone properties was largely unexploited at the time [204], while yet another study even concluded that US assessment should not yet be considered a valid replacement for radiographic bone age determination [205]. Later, the QUS device SonicBone™ was found to be safe, convenient and non-painful while yielding highly reproducible results [206]. The same study stated that the BoneAge™ system was not commonly used due to its lack of consistency and confirmatory data in large groups and claimed that the newer SonicBone™ device took most of the drawbacks of both the Greulich-Pyle atlas and previously suggested QUS based devices into consideration [206]. A successor to the SonicBone™ device is the BAUS™ system, which also performs assessments of bone age using US. Its results have been found to be highly reproducible and comparable to bone age assessments based on radiographs [27]. The differences between the determined bone ages of BAUS™ and assessments of Greulich-Pyle and TW3 atlases were found to be non-significant. However, two different parameters from three different body locations had to be taken into account to achieve this result.

All previously mentioned systems determine absolute bone ages, are mainly used to analyze and predict growth in children and are unsuitable to determine the growth plate ossification. Work published in 2021 examined “the left wrist of 688 (322 males, 366 females) patients between the ages of 9 and 25 years” by ultrasonography. This study found that “the data obtained may help determine legally critical age limits of 14 and 15. Although it does not seem useful for the age of 18, ultrasonography may be recommended in selected cases as a fast, inexpensive, frequently reproducible radiological method without concern about radiation and without a predictable health risk” [207].

Recently, a multi-channel CNN combined with a sliding window scheme based on 1-D QUS signals was devised as an osteoporosis diagnosis method. The data used for that publication was collected at 1/3 of the distal radius of the non-dominant hand using an US bone sonometer [208]. Even though that work does not yield any information concerning the bone age of subjects, it vividly demonstrates the capabilities of QUS based approaches.

4.3. Materials and methods

This section covers the details of the developed mobile system in Section 4.3.1. Section 4.3.2 provides an overview of the conducted corresponding medical study, while Section 4.3.3 explains the underlying measurement principle applied in this study. Section 4.3.4 illustrates the data processing pipeline used to classify the signals.

4.3.1. System design

An integral part of the developed mobile system are custom US acquisition electronics built by *Fraunhofer Institute for Biomedical Engineering (IBMT)*. These electronics were also used to acquire muscle contraction and muscle fatigue state signals (see Section 3.3.1.1) and consist of two US array transducers with four elements in total which are used to measure transmission and reflection signals. Each element has an aperture dimension of 11x11mm, a center frequency of 1.0 MHz and can be accessed individually for measurements. The transducers are built out of custom 1-3 piezo composite materials and have a natural focus at a depth of 1.5 cm inside the radius or ulna bone. The acoustic pressure distribution of this natural focus has been designed to provide optimal growth plate detection possibilities for a human wrist. All components are integrated in an electromagnetically shielded housing and the acoustic output was evaluated according to medical standards in collaboration with certified laboratories. Figure 4.1 shows the complete portable integrated US-based measurement device. Figure 4.2 displays the compact, stackable US electronic module used in the device (a) and the inbuilt US array (b), while Figure 4.3 presents the simulated distribution of the acoustic pressure of a single array element for a sound speed value of ≈ 2500 m/s in bone tissue.

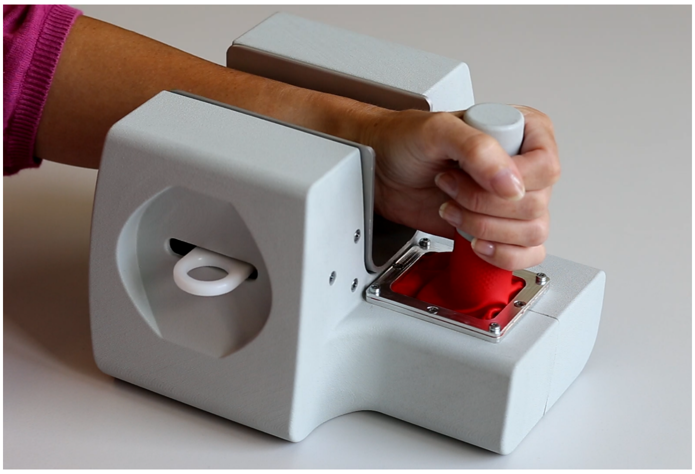


Figure 4.1.: The portable integrated US-based measurement device (taken from [1]).

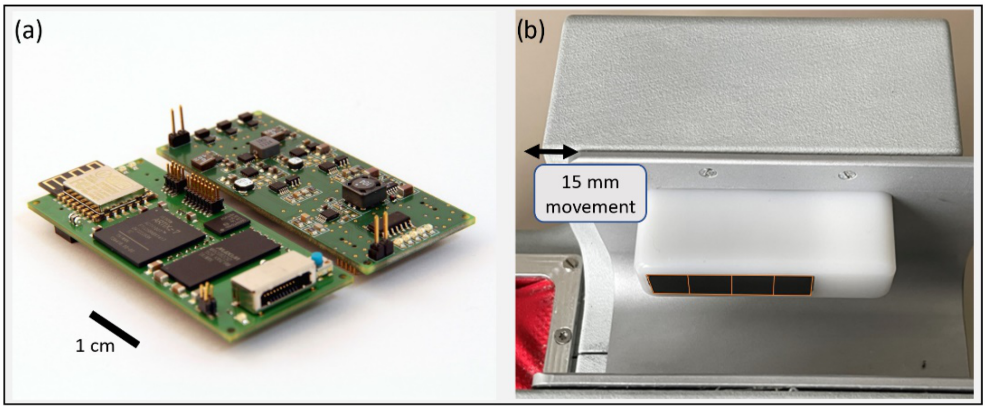


Figure 4.2.: The compact, stackable US electronic module used in the device (a) and the inbuilt US array (b) (taken from [1]).

4.3.2. Study design

A clinical study containing data from 148 girls and women in total was conducted at the *Saarland University Medical Center* and approved by the medical association of Saarland (ID: 255/15). This study included patients from paediatric endocrinology and healthy volunteers. The portable system described in Section 4.3.1 was an integral part for the data acquisition of this study. An experienced paediatric endocrinology consultant, blinded to the age, height and weight of each patient, determined the bone age from radiographs using the Greulich and Pyle bone age atlas (see Section 4.1.2) of patients with a clinical indication. Five subjects, whose arm physiques were unsuitable for the device, were excluded from the study. Those patients either had wrists that were too slim, causing acoustic coupling issues, or too large, such that the engine powering the US probe was unable to move along the forearm. No subject reported any discomfort as the top skin layer was being pulled slightly (within a 15 mm range) and the US transducer array was being pushed towards the forearm. Subjects with known medical

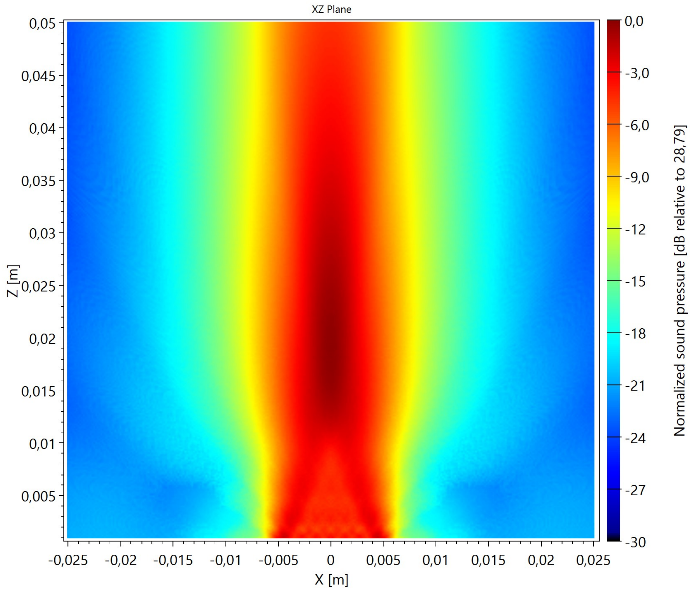


Figure 4.3.: Simulation of the acoustic pressure distribution of a single array element for a sound speed value of ≈ 2500 m/s in bone tissue (taken from [1]).

conditions affecting the bone age were excluded when radiographs were not available. Bone ages derived from radiographs or chronological ages were used as annotation ground truth to distinguish between open and ossified growth plates. Figure 4.4 shows a flowchart of the examined cohort in accordance with the *Standards for Reporting of Diagnostic Accuracy Studies (STARD)* guidelines. Reasons for the exclusion of potentially eligible subjects were previous fractures of the left wrist, hand, or underarm; precocious puberty; pubertas tarda; growth hormone deficiency; congenital adrenal hyperplasia; previous radio- or chemotherapy; previous high-dose steroid therapy; Turner syndrome; Down syndrome; transgender individuals undergoing/previous hormone therapy; tall stature individuals undergoing hormone therapy; non-treated hyper- or hypothyroidism or severe obesity (BMI > 40). 120 subjects remained in the study after exclusions.

The data acquisition workflow consisted of eight consecutive signal acquisition intervals (i.e. one for each transducer), yielding reflection and transmission signals for each subject. In total, less than one minute per subject was required for the necessary four back-and-forth motions of the motorized sled.

4.3.3. Measurement principle

The complete system features a grab handle and a fixture containing US transducers (see Figure 4.1). Two US array transducers with four elements each gently pressed against the arm laterally and medially, respectively. Both US transducer arrays on each side of the arm acquire pulse-echo reflection signals and receiving transmission signals. The system moves all US transducers in distal and proximal directions during data acquisition to increase the chances of measuring through the potentially open growth plate.

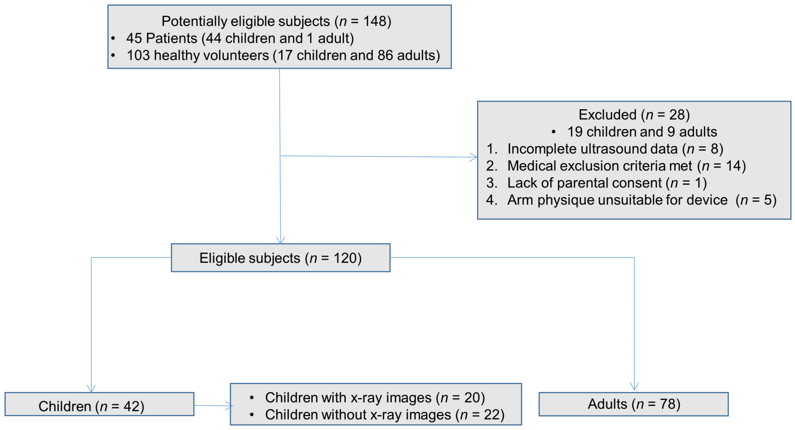

Potentially eligible subjects (*n* = 148)
• 45 Patients (44 children and 1 adult)
• 103 healthy volunteers (17 children and 86 adults)
Excluded (*n* = 28)
• 19 children and 9 adults
1. Incomplete ultrasound data (*n* = 8)
2. Medical exclusion criteria met (*n* = 14)
3. Lack of parental consent (*n* = 1)
4. Arm physique unsuitable for device (*n* = 5)
Eligible subjects (*n* = 120)
Children (*n* = 42)
• Children with x-ray images (*n* = 20)
• Children without x-ray images (*n* = 22)
Adults (*n* = 78)

Figure 4.4.: STARD 2015 flowchart of the study cohort (taken from [1]).

The use of US array transducers with few elements minimizes active movements of the subject as the skin's elasticity permits movements of up to 15 mm without having to reposition any transducers. The grab handle restrains the subject's arm movements and prevents any involuntary actions. It also serves as a reference point for the arm, reducing the variation of potential measurement locations. The linear orientation of the US transducer elements enables measurement locations of subjects with different arm lengths to be taken into account. The system was designed for forearm sizes ranging from 31 mm to 55 mm at the epiphyseal plate of the radius bone, which was sufficient in most cases. The system does not perform 2-D US imaging and acquires only raw 1-D A-scans. Rectangular tri-state burst excitation signals with a length of 3 or 5 periods at center frequency were used for each measurement path. In each measurement modality, 25 A-scans containing $2,048$ values each were sampled as a motorized sled moved the transducers along the forearm (see Figure 4.2 (b)). Figure 4.3 shows the simulated acoustic pressure distribution of a single array element for a speed of sound value of approximately 2500 m/s in bone tissue, while Figure 4.5 depicts a sample reflection and transmission signal. The signal sections highlighted in orange boxes were the most significant ones as they presumably contained the most information on the underlying tissue and potential bone structures. The sections at the beginning and the end of the signal were usually artifacts resulting from deep or very shallow soft tissue structures.

4.3.3.1. Choice of frequencies

The system works at an US center frequency of 1 MHz. In related literature, frequencies below 1.5 MHz were often used for axial transmissions, in which an US transmitter/receiver pair was placed in parallel to the bone's longitudinal axis on one side. A frequency of 0.75 MHz has been previously used for transmitting US waves through the left wrist [202]. A higher frequency of 3 MHz has been used in measurements of axial transmission in the thin cortex

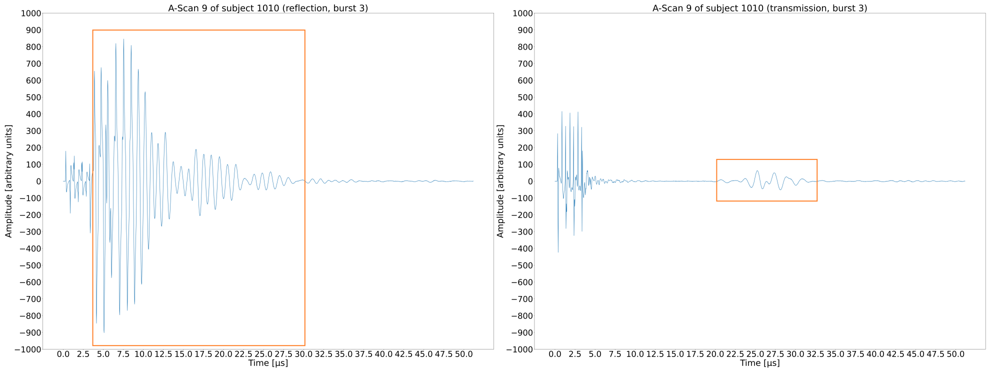


Figure 4.5.: A-scan samples illustrating the signals stemming from reflection (left) and transmission (right) of the most distal element(s) acquired with a 3-cycle burst. The orange boxes indicate signal parts containing the most significant information (taken from [1]).

of human long bones [209]. The thickness of cortical bone was in the range of several millimeters [209]. As the speed of US wave propagation in human bones depends on a variety of influencing factors, with the most prominent being age and sex [210], there was a need to find a compromise enabling valid measurements in a large age range in children and adolescents. Hence, a value of 2500 m/s for the simulations performed before assembling the system was assumed (see Figure 4.3).

4.3.3.2. Tissue-dependent ultrasound wave behavior

A growth plate without a full ossification leads to comparatively slower US waves as they travel through more non-bony soft tissue and less-ossified bone tissue. Due to the lateral extension of the acoustic beam, some US waves travel through bone tissue located outside the growth plate. This can result in an overlay of multiple US wave traveling paths, with the signal passing through the open growth plate merging with signals stemming from other tissue parts. Furthermore, the exact location of a potentially open growth plate varies from subject to subject. Hence, the signal analysis has to be able to detect signal differences caused by changes in relative orientation of the bone and the US transducers during transducer movements. In subjects with an open growth plate, the measurements may differ more significantly from the signals measured next to the epiphyseal gap, through or around the bone tissue. This can be caused by less attenuation and change in US propagation speed due to different stages of ossification or changes in bone surface structure that reflect and diffract the US wave at the boundaries of the growth plate. The presence of an open growth plate potentially leads to larger changes in reflectivity or in the ratio of transmitted US waves as the reflectivity and intensity of transmitted US waves differ more than the constantly changing signals moving over an ossified growth plate (see Figure 4.6). The deployed algorithms aim to distinguish between those two different kinds of signals.

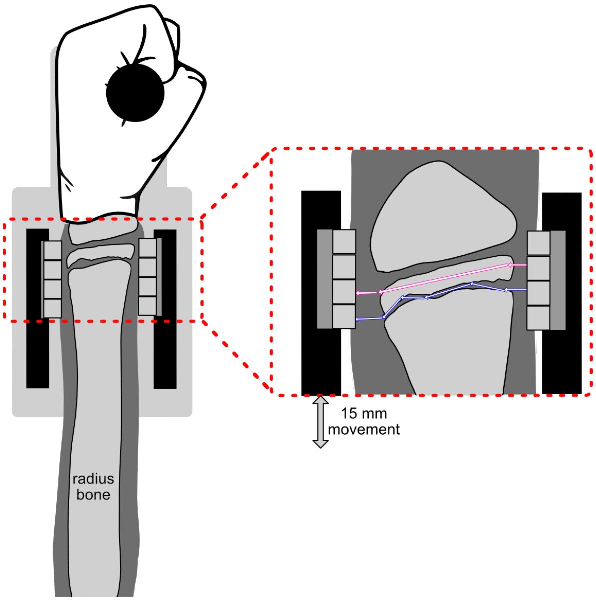
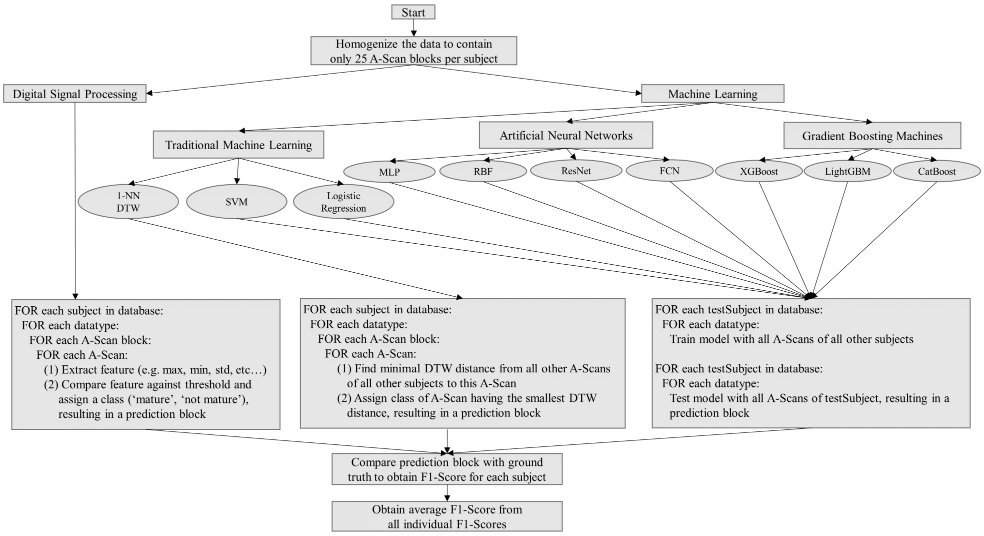


Figure 4.6.: The left forearm was placed in the device with the transducers of the motorized sled positioned around the assumed location of the growth plate (left). The acoustic wave propagation through the growth plate (blue) and the radius bone (pink) is depicted in the enlarged detail inset on the right (taken from [1]).

4.3.4. Data processing pipeline

Figure 4.7 shows the general workflow of the complete data processing pipeline. Firstly, the input data was homogenized after importing it to ensure that


Start
Homogenize the data to contain only 25 A-Scan blocks per subject
Digital Signal Processing
Machine Learning
Traditional Machine Learning
Artificial Neural Networks
Gradient Boosting Machines
MLP
RBF
ResNet
FCN
XGBoost
LightGBM
CatBoost
1-NN DTW
SVM
Logistic Regression
Compare prediction block with ground truth to obtain F1-Score for each subject
Obtain average F1-Score from all individual F1-Scores

```
FOR each subject in database:
  FOR each datatype:
    FOR each A-Scan block:
      FOR each A-Scan:
        (1) Extract feature (e.g. max, min, std, etc…)
        (2) Compare feature against threshold and
            assign a class (‘mature’, ‘not mature’),
            resulting in a prediction block
```

```
FOR each subject in database:
  FOR each datatype:
    FOR each A-Scan block:
      FOR each A-Scan:
        (1) Find minimal DTW distance from all other A-Scans
            of all other subjects to this A-Scan
        (2) Assign class of A-Scan having the smallest DTW
            distance, resulting in a prediction block
```

```
FOR each testSubject in database:
  FOR each datatype:
    Train model with all A-Scans of all other subjects

FOR each testSubject in database:
  FOR each datatype:
    Test model with all A-Scans of testSubject, resulting in a
    prediction block
```

Figure 4.7.: General workflow of the complete data processing pipeline (taken from [1]).

only 25 A-scans blocks existed per subject. This was a necessary step as an inhomogeneous amount of A-scans block existed before for each subject. After this step, the data was further processed either by DSP or ML methods. The aim of all methods was to obtain binary predictions by grouping all subjects with bone ages <18 years and bones ages ≥18 years into different categories.

4.3.4.1. Digital signal processing

If the data was processed by DSP, a naïve threshold algorithm performed a binary classification of the signals by extracting a certain feature (e.g. maximum value, minimum value, standard deviation, etc.) from each A-scan of each A-Scan block. The algorithm performed this action for each possible type of signals (such as reflection or transmission) and for each subject of the complete database. Comparing the extracted features to a threshold resulted in a binary prediction for each block. In total, this process resulted in 25 predictions for a single block of each subject. The F_1 -Score for each subject was computed by comparing the predictions to the ground truth. Averaging all F_1-scores of all subjects lead to the final F_1-Score, which was the metric for the performance of the algorithm.

4.3.4.2. Machine Learning

If the data was processed by ML, the corresponding algorithms performed a binary classification by using each A-scan of each A-Scan block that does *not* belong to the currently observed subject as training input. All A-scans from the A-Scan block of the currently observed subject were temporarily withheld to be later used as testing signals for evaluation. This protocol enforced a strict LOOCV regime. The algorithm performed this action for each possible type of signals (such as reflection or transmission) and for each subject. If the 1-NN DTW algorithm was deployed, the algorithm performed a binary classification of the signals by finding the minimal dynamic time warping distance of each A-Scan of each A-Scan block to any other A-Scan. The algorithm performed this action for each possible type of signals (such as reflection or transmission) and for each subject. In total, this process resulted in 25 predictions for a single block of each subject. The F_1 -Score for each subject was computed by comparing the predictions to the ground truth. Averaging all F_1-scores of all subjects leads to the final F_1-Score, which is the metric for the performance of the algorithm.

4.4. Results

Section 4.4.1 presents the results of DRTs that have been applied on the input data, while Section 4.4.2 summarizes the results of all used ML methods.

4.4.1. Dimensionality Reduction Techniques

Section 4.4.1.1 contains the results of a 2-D and a 3-D PCA of the input data and Section 4.4.1.2 contains the results of a 2-D t-SNE visualization of the input data. In all scenarios, the underlying data were signals obtained with a burst of 3 cycles on transmission signals of transducer number 3.

4.4.1.1. Principal Component Analysis

Figure 4.8 and Figure 4.9 show a 2-D and a 3-D PCA (see Section 2.7.3.1) visualization of the input data, respectively. In both cases, green dots represent A-scans belonging to subjects below the age of 17 years. Yellow dots represent A-scans belonging to subjects between the age of 17 years and 19 years and red dots represent A-Scans belonging to subjects older than 19 years.

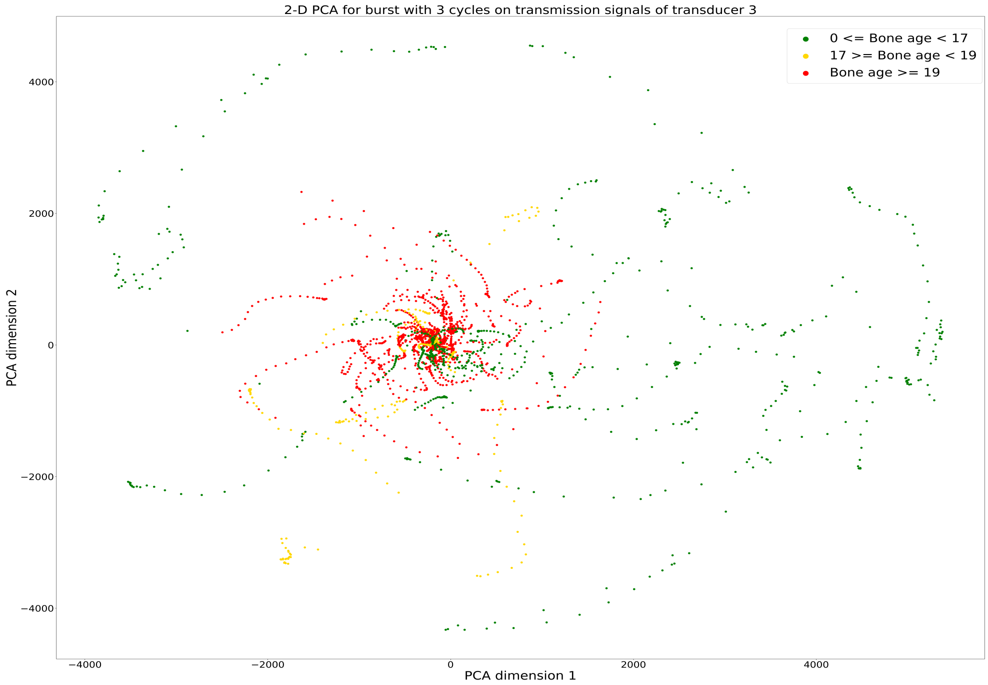


Figure 4.8.: 2-D PCA visualisation showing the interrelation of all A-scans. Color codes for bone age groups: green = <17 years, yellow = 17 years to <19 years and red = ≥19 years.

In both PCA figures, there is no strict separation of signals but a clear tendency of signals belonging to the age group ≥19 years to group together with most signals belonging to the age group 17 years to <19 years located close to this gravitational center. Signals belonging to the age group < 17 years tend to be located further away from this center. This grouping tendency substantiates the main hypothesis that signal classification in general is possible. Figure 4.10 shows the ratio of explained variance of the first 10 principal components of the data. The first two principal components combined contain ≈ 57.20 % of the explained variance, while the first three principal components combined contain ≈ 68.01 % of the explained variance. Once again, this observation substantiates the main hypothesis as a significant amount of explained variance is clearly distinguishable.

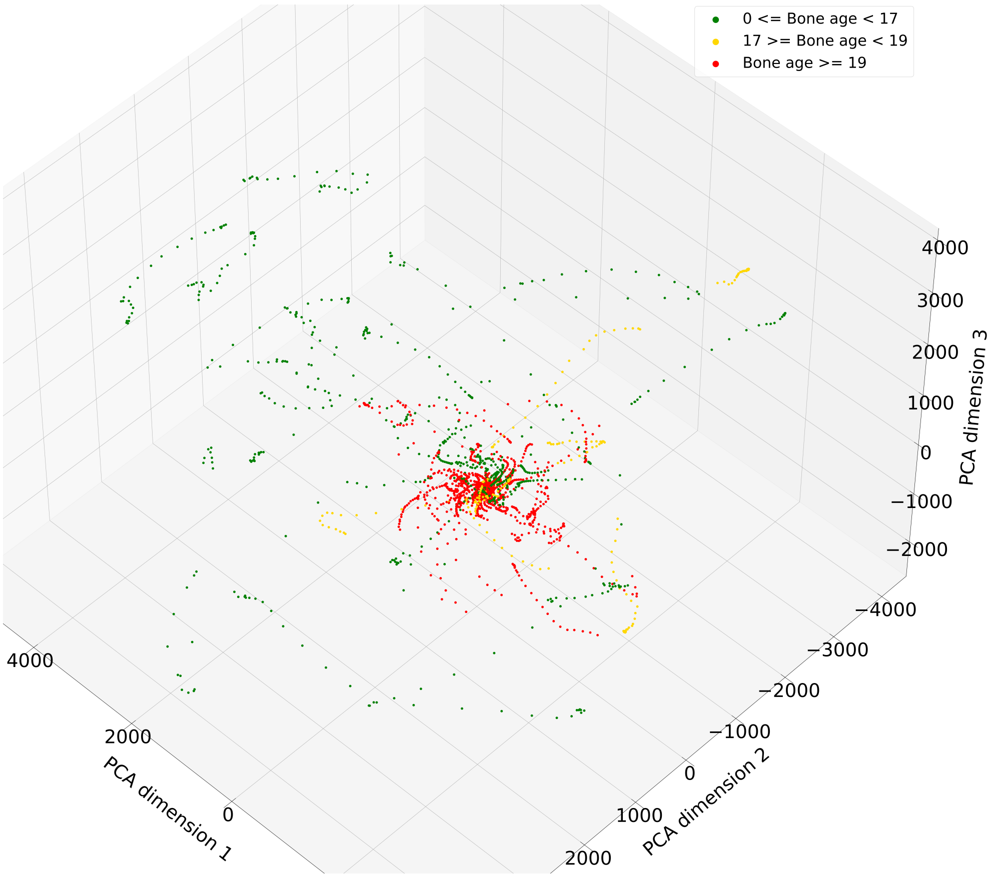


Figure 4.9.: 3-D PCA visualisation showing the interrelation of all A-scans. Color codes for bone age groups: green = <17 years, yellow = 17 years to <19 years and red = ≥19 years (adapted from [1]).

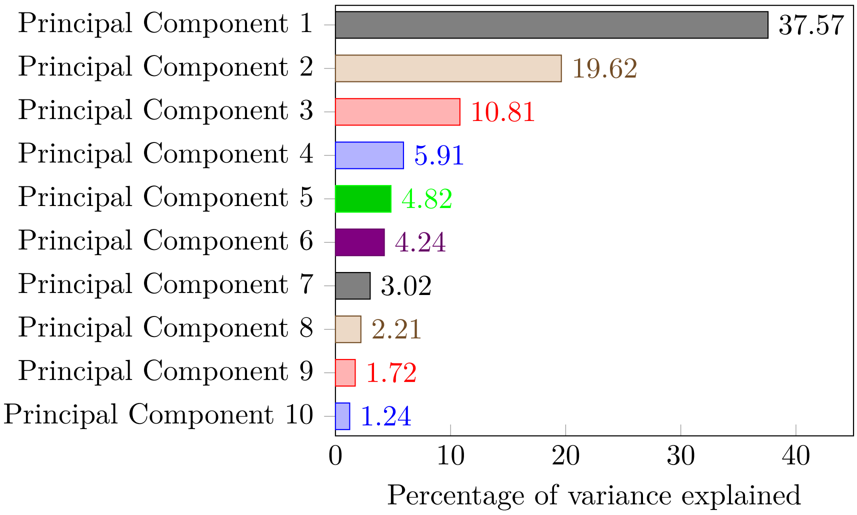


Figure 4.10.: Ratio of explained variance of the first 10 principal components of the data (taken from [1]).

4.4.1.2. t-distributed stochastic neighbor embedding

Figure 4.11 shows a t-SNE visualization of the input data (see Section 2.7.3.2. Green dots represent A-scans belonging to subjects below the age of 17 years. Yellow dots represent A-scans belonging to subjects between the age of 17 years and 19 years and red dots represent A-Scans belonging to subjects older than 19 years.

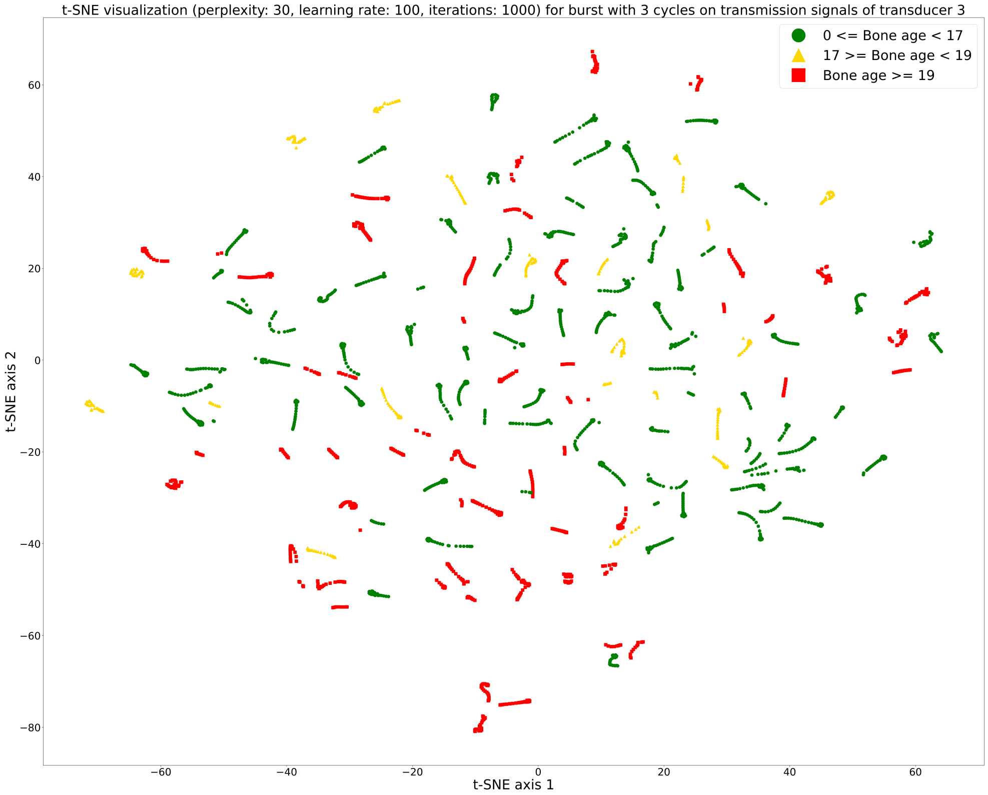


Figure 4.11.: 2-D t-SNE visualisation showing the interrelation of all A-scans. Color codes for bone age groups: green = <17 years, yellow = 17 to <19 years and red = ≥19 years (taken from [1]).

4.4.2. Machine Learning

The DRT results (see Section 4.4.1) indicate that a categorization of the input signals should be possible. To confirm this hypothesis, this section describes the results of a comprehensive analysis including various ML techniques, that used the raw, unprocessed 1-D US signals as input data. All signals were acquired with a burst of 3 cycles on transmission signals of transducer number 3. Table 4.1 summarizes the average F_1-scores of various models for different age groups. Additionally, it also provides the time for training and evaluation. In each column the best and the second best achieved performances are highlighted in green or yellow, respectively. Appendix B presents the detailed results for each subject and each examined model.

Model	Average F_1-score for all subjects (%)	Average F_1-score for all subjects under 17 years (%)	Average F_1-score for all subjects under 18 years (%)	Average F_1-score for all subjects between 17 and 19 years (%)	Average F_1-score for all subjects over 18 years (%)	Average F_1-score for all subjects over 19 years (%)	Time for training and evaluation (h)
CatBoost (10,000 iterations)	86.93	83.68	78.48	82.57	90.72	91.35	26.08
CatBoost (1,000 iterations)	86.40	76.38	91.30	81.19	83.86	91.20	2.65
XGBoost	85.23	80.22	76.67	78.86	90.61	91.85	0.08
CatBoost (100 iterations)	84.77	77.24	88.41	81.84	79.43	89.45	0.23
LightGBM	82.57	74.70	72.95	77.71	90.38	90.33	0.06
FCN	82.30	77.62	72.29	70.00	89.80	91.71	186.18
1-NN DTW	81.50	65.62	64.86	75.00	91.54	91.42	4.31
ResNet	78.53	61.84	59.43	74.29	91.19	91.93	271.75
MLP	77.73	66.49	63.81	66.00	86.14	85.75	47.99
SVM	77.00	70.59	66.95	67.43	82.97	85.86	0.18
Logistic Regression	75.20	61.51	59.71	71.00	85.74	86.55	0.82
Radial Basis Functions Neural Network	73.10	50.03	47.33	78.71	88.0	85.60	0.23

Table 4.1.: Summary of the average F_1-scores of various models for different age groups. In each column the best and the second best achieved performances are highlighted in green or yellow, respectively.

4.5. Discussion of epiphyseal plate detection

This section closes the chapter by discussing the applied DRTs and ML methods in Section 4.5.1 and Section 4.5.2. Section 4.5.3 provides an overview of the limitations of the chosen approach, while Section 4.5.5 concludes the discussion.

4.5.1. Dimensionality Reduction Techniques

With the help of DRTs, a better understanding of the underlying data is possible. The 2-D PCA (see Figure 4.8) and the 3-D PCA (see Figure 4.9) both show that the A-scans clustered strongly according to age segments. Even though there are some overlapping data-points, the overall distribution indicates the data-points belonging to the age segment >19 years cluster more densely than the data-points belonging to other age segments. The 2-D t-SNE visualization (see Figure 4.11) illustrates that the A-scans tend to cluster according to certain age segments but cluster for each subject much more densely. Even though neither DRT shows a clean visual segregation between A-scans belonging to different categories, the percentages of explained variances devised from the PCA (see Figure 4.10) indicate that the A-scans contain a significant amount of information in comparison to noise. This finding encouraged a deeper analysis and classification of the signals by deploying several ML algorithms.

4.5.2. Machine Learning

A classic naïve DSP threshold algorithm did not achieve a satisfactory signal classification performance. ML methods, on the other hand, managed to yield much better results. GBMs outperformed ANNs in terms of accuracy and speed for almost all examined age segments. These results are in line with expectations as MLPs, FCNs and ResNet each exhibit significant drawbacks as discussed in Section 2.7.4.6. However, ResNet still performs comparatively well for the bone age segment >18 years, while very shallow RBFNNs achieved the lowest performance in total. The fastest models, LightGBM and XGBoost, yielded results within minutes, while the slowest model, ResNet, needed more than 11 days to complete. The best total average F_1-score of ≈ 87 %, obtained using a CatBoost model, showed that robust binary classifications of the examined signals are possible. It is especially noteworthy that XGBoost yielded an average F_1-Score of 85.23 % after approximately 5 minutes only, while CatBoost achieved the best average F_1-Score of 86.93 % after approximately 26 hours. This shows that near real-time classifications of the signals are possible if minimal performance penalties are deliberately taken into account.

In the bone age segment >18 years, all models performed better compared to bone age segments containing only younger subjects. The premature closing of the growth plate or artifacts from stiffer soft tissue regions might be responsible for misclassifications in some younger subjects. F_1-scores of around 80 % for bone age segments with younger subjects and F_1-scores of more than 90 % for bone age segments with older subjects provide a reasonable indicator, although further refinements are needed if the device is to be deployed for medical purposes or in general clinical usage.

4.5.3. Limitations

A potential limitation of the conducted clinical study is its dependence on data stemming from girls and women of a single comparatively homogenous and predominantly white population of the federal state of Saarland, situated in south-western Germany. This population may not be representative of other international populations and the presented models may not be applicable for all ethnicities and sexes [200, 201].

4.5.4. Future work

The relative homogeneity of the examined population, as described above, warrants further research. For example, multi-centre studies that also include male and ethnically diverse subjects. More sophisticated ML methods, such as Transformers (see Section 2.7.4.6), remain to be explored in the future. Furthermore, the effect of mathematical transformations or filters, described in Section 2.3.2, on various ML methods should also be explored.

Appendix B presents the F_1-scores for each ML model and each study subject. The careful examination of Table B.1 shows that signal classifications of certain subjects always yield rather low F_1 scores, regardless of the chosen ML model. Further efforts should be made to investigate the nature of these phenomena and whether better results could be achieved if certain signals were excluded from the experiments. Erroneous signals or artifacts might have occurred due to technical issues, such as the absence of ultrasound gel on certain arm areas, faulty measurements or incomplete acquisitions.

4.5.5. Conclusion

The presented approach uses ML methods to classify 1-D US raw signals in order to determine ossification states of the distal growth plate of the radius and ulna bones of girls and women. Using 1-D US signals to categorize the distal growth plate as being either in an open or an ossified state, the system was also able to detect the completion of bone growth, which "stops around the age of 21 for males and the age of 18 for females when the epiphyses and diaphysis have fused (epiphyseal plate closure)" [211]. This non-invasive approach is not based on ionizing radiation, does not require any DSK of specialized medical practitioners and provides a viable alternative to existing radiography-based methods like X-ray imaging for the determination of completed bone growths.

Chapter 5

Identification of hepatic steatosis and fibrosis in patients with non-alcoholic fatty liver disease

Liver diseases are an issue of growing global concern. *Liver fibrosis*, the excessive accumulation of extracellular matrix proteins, and *liver steatosis*, a condition where excess fat builds up in the liver, often accompany liver diseases. Transient Elastography (TE) is often deployed as a non-invasive method to assess liver fibrosis or steatosis stages but the necessary equipment for this technique is comparatively complex and expensive. This chapter shows how ML models based on 1-D US signals, that can be acquired in a non-invasive fashion, can be used for the assessment of different stages or grades of liver fibrosis and liver steatosis. Partial results of the presented work in this chapter have been published in [6].

Contents

5.1. Introduction

Chronic liver diseases are a "major cause of morbidity and mortality worldwide" and "has a varied aetiology, including viruses, such as hepatitis B (HBV) and hepatitis C (HCV). Worldwide, over half a billion people may be chronically infected with either of these viruses. Metabolic causes include the increasing prevalence of [NAFLD]. Toxic causes, such as excess alcohol consumption, aflatoxin exposure and autoimmune disorders, such as primary biliary cirrhosis and autoimmune hepatitis, [also] contribute to the disease burden" [212]. NAFLD is characterized by excessive liver fat depositions while other liver disease aetiologies, such as alcohol-related liver disease or chronic viral hepatitis are absent. Nonalcoholic steatohepatitis (NASH) is the progressive and inflammatory form of NAFLD and can potentially progress to an advanced liver disease stage, cirrhosis or hepatocellular carcinoma. The amount of people suffering from NAFLD is "constantly increasing" and its global prevalence is "currently estimated to be 24 %". A high NAFLD prevelance "with potential for progressive liver disease creates challenges for screening, as the diagnosis of NASH [currently] necessitates invasive liver biopsy" [213]. *Liver fibrosis* and *liver steatosis* are well recognized accessory phenomenons of chronic liver diseases [212]. This chapter presents an approach to classify liver fibrosis and steatosis stages based on 1-D US signals of patients suffering from NAFLD. Section 5.1.1 discusses the importance of *grading* and *staging* in detail. Section 5.1.2 and Section 5.1.3 provide an overview of liver fibrosis and steatosis.

5.1.1. Grading and Staging

The disease *stage* indicates how far the disease has already progressed, with the end stage resulting in death of the patient or organ failure. The disease *grade* reflects how quickly the disease is progressing towards the end stage. In most chronic liver diseases, the end stage is *cirrhosis* with clinical decompensation, whereas earlier stages have smaller degrees of fibrosis or cirrhosis (see Section 5.1.2). The grade is a measure of severity of the underlying disease, with disease properties that vary with type and pattern of injury. Ideally, both grade and stage should predict prognosis and guide therapeutic intervention [214].

5.1.2. Liver Fibrosis

Liver fibrosis is an abnormal wound repair response in the liver caused by a variety of chronic injuries, which is characterized by over-deposition of diffuse extracellular matrix (ECM) and anomalous hyperplasia of connective tissue. Liver fibrosis may further develop into liver cirrhosis, liver cancer or liver failure. It is known that early liver fibrosis stages are reversible but the detailed mechanisms behind the decrease of fibrotic scarring are still unclear.

Several grading and staging systems for the liver fibrosis assessments exist with the first major scoring system developed as early as 1981. Later, the *Metavir* scoring system was developed to specifically evaluate hepatitis C-related liver diseases. In this system, the fibrosis stage is assigned a scored within the rage from 0 to 4. Table 5.1 summarizes the Metavir scoring system [214]:

Fibrosis stage	Description
F0	No fibrosis
F1	Mild fibrosis
F2	Moderate fibrosis
F3	Severe fibrosis
F4	Cirrhosis

Table 5.1.: Summary of the Metavir system for liver fibrosis staging.

5.1.3. Liver Steatosis

Liver steatosis is also known as fatty liver disease and defined "as the presence of large and small vesicles of fat, predominantly triglycerides, accumulating within hepatocytes". Hepatic steatosis is "frequently associated with *obesity*, *insulin resistance* and *dyslipidaemia*" and "may also be a result of secondary causes, which include *alcoholism*, [*hepatitis C*], *severe weight loss*, *total parenteral nutrition* or [*certain drugs*]". Several scoring systems for liver steatosis exist to quantify liver fat. "The grade of steatosis is based on the proportion of hepatocytes containing visible lipid and is expressed" as described in Table 5.2 [212].

Steatosis grade	Proportion of hepatocytes containing visible lipid
0	<5 %
1	5 % - 33 %
2	33 % - 66 %
3	>66 %

Table 5.2.: Steatosis grades based on the amount of hepatocytes containing visible lipid.

5.2. State of the art

Liver disease diagnoses are complex and several approaches have been proposed to develop robust and accurate methods in the past. The most traditional method is *liver biopsy* (see Section 5.2.1). As this approach has many disadvantages and has been described as not being “a viable tool for widespread NAFLD risk stratification” [215], other non-invasive methods have been developed. Section 5.2.2 describes US based methods, while Section 5.2.3 covers alternative methods relying on other medical imaging technologies. Section 5.2.4 compares and evaluates those approaches w.r.t possible mobile low-cost solutions.

5.2.1. Liver biopsy

For decades, liver biopsy has been the gold standard for assessing a variety of different diseases, especially in the field of chronic liver disease diagnosis even though this procedure has “an associated but low morbidity and mortality risk” [212]. Liver biopsy serves two principal functions: It establishes or confirms the diagnosis of a particular type of liver disease and it is also frequently used to assess the severity of the disease. For a skilled interpreter, a biopsy is appropriate for the establishment or confirmation of a diagnosis, which is called *qualitative analysis*. However, it is less reliable for the assessment of disease severity, which is called *semiquantitative analysis* [214]. Overall, liver biopsies suffer from the following limitations [212]:

1. **Morbidity and mortality**: “Percutaneous liver biopsy has a small, but quantifiable, risk of mortality, quoted as between 1 in 1000 and 1 in 10000 patients”.
2. **Complications**: “Post-procedural pain or a localised haematoma occur in between 3 % and 30 % of cases. More severe complications [...] may occur in 0.3 % to 0.6 % of cases. [Liver] tumor seeding following biopsy of suspected carcinomas in cirrhosis may also occur in 2.7 % of cases”.
3. **Sampling variability**: “Percutaneous liver biopsy typically samples less than 1/50000th of the liver, so any heterogeneity of pathological features may lead to sampling variety”.
4. **Subjectivity and inter-observer variation**: “Histological scoring systems are designed to be objective and reproducible but interpretation [by humans] is still a source of error. Inter-observer variability is low for the assessment of fibrosis, but higher for the assessment of activity or inflammation”.

5.2.2. Ultrasound based methods

5.2.2.1. Acoustic Radiation Force Impulse

Acoustic Radiation Force Impulse® (ARFI) imaging is an US technique measuring soft tissue displacements by inducing acoustic pulses with high energies in the soft tissue to be examined. The displacements induced by these pulses are “quantified and interpreted as a measurement of liver stiffness”. It has been shown that ARFI based techniques are very accurate for the detection of very pronounced or severe fibrosis and cirrhosis. ARFI is also comparable to TE (see Section 5.2.2.2) for the assessment of very pronounced fibrosis and cirrhosis and “was significantly more likely to obtain reliable measurements” for the same patients. However, its suitability for “detecting earlier stages of [liver] fibrosis [remains] limited” [212].

5.2.2.2. Transient Elastography

TE is an US based technique, which was first developed in 1992. “It requires a [1-D] probe and an [US] transducer. The probe transmits a low-amplitude mechanical pulse to the liver, “inducing the propagation of an elastic shear wave”, which propagates through the soft tissue. “The propagation velocity, measured by the transducers, is positively related to the liver stiffness”. This method is an integral part of the *Fibroscan* device, which became the first U.S. Food and Drug Administration (FDA) approved US based elastography technique in 2013 [216]. TE has “a very high sensitivity and specificity for detecting cirrhosis” but its accuracy has been shown to be lower for the detection of earlier fibrosis stages. *Fibroscan* machines “measure hepatic steatosis levels by using [a parameter called Controlled Attenuation Parameter® (CAP)], which demonstrated an “excellent” correlation between different steatosis grades [212]. In general, it has been shown that TE is “useful for the staging of liver fibrosis in patients with NAFLD, particularly for those with advanced fibrosis and cirrhosis” [216].

5.2.2.3. Shear Wave Elastography

2-D Shear wave elastography (SWE) is a method that was first commercially deployed in a diagnostic imaging device called *Aixplorer®*. For liver disease diagnoses, 2-D SWE focuses acoustic radiation force impulses to “multiple sites in the liver and generates [a] real-time [...] 2-D quantitative [map] of liver tissue elasticity” [...] over a significantly larger tissue area [...] than [TE] (see Section 5.2.2.2) and [ARFI] (see Section 5.2.2.1). The measured speed of the induced shear waves can be converted to a numerical value representing the soft tissue stiffness in Kilopascal (kPa). A study published in 2020 mentioned that “no meta-analyses of studies of 2-D SWE in patients with NAFLD have been conducted. Therefore, the diagnostic accuracy of 2-D SWE in patients with NAFLD requires further investigation” [216].

5.2.2.4. B-Mode ultrasound

Classic *B-Mode US* is “widely used clinically to detect hepatic steatosis, but it can also detect the vascular changes of chronic liver disease with contrast enhancement”. B-Mode US “is also able to detects hepatic steatosis, based on the premise that steatosis causes increased echogenicity of the hepatic parenchyma, leading to a brighter image when compared to the cortex of the ipsilateral kidney". Certain conditions, such as obesity reduce the accuracy of B-Mode US because of the increased attenuation of signals caused by subcutaneous fat. This technique also performs "more poorly for the quantification of hepatic lipid" [212]. However, B-Mode US is still the most frequently used primary imaging modality for the evaluation of liver diseases. A recent publication from 2020 found modern high-end US devices to be "an excellent method to detect advanced steatosis in patients with various chronic liver diseases". Even though those devices might be suitable for the diagnosis of mild steatosis, they have a higher sensitivity at the expense of specificity. The stages of fibrosis and etiology of chronic liver diseases seem not to impact the diagnostic accuracy [217].

5.2.2.5. Quantitative ultrasound

Attenuation and *backscatter* coefficients deduced from QUS measurements have been deployed for liver fat quantification and to distinguish between the steatosis grades 1, 2 and 3. The attenuation coefficient measures US energy loss in soft tissues and "provides a quantitative parameter analogous to the qualitative loss of view of deeper structures observed in severe fatty [livers]". The backscatter coefficient "measures the returned [US] energy from [soft] tissue and provides a quantitative parameter analogous to echogenicity". Backscatter coefficients showed an "excellent diagnostic performance for quantification of hepatic steatosis compared to Magnetic resonance imaging proton density fat fraction (MRI-PDFF)" (see Section 5.2.3.5). Although QUS parameters might potentially be usable to quantify liver steatosis accurately, further "assessment of variation across different scanner manufacturers and operators is [necessary] to further investigate accuracy, reproducibility and repeatability" of the examined techniques [215].

5.2.2.6. A-Mode ultrasound

1-D US signals from the liver contain rich information about liver microstructure and composition. A study that was conducted on 204 subjects and published in 2020 found that DL algorithms using these signals "are accurate for [the] diagnosis of nonalcoholic fatty liver disease and hepatic fat fraction quantification when other causes of steatosis are excluded" [25]. In this study, two 1-D CNN algorithms were developed: a binary classifier and a fat fraction estimator. The former distinguished between subjects with and without NAFLD, while the latter output the predicted fat fraction as a percentage [25].

5.2.3. Alternative Methods

5.2.3.1. Computed tomography

Computed Tomography (CT) is used to assess hepatic steatosis based on radiographic densities. "Unenhanced CT is more specific than US for NAFLD detection" and dual-energy CT has been shown to more accurately "quantify hepatic steatosis and potentially permit fibrosis staging". It also has the potential to quantify liver fat contents [215]. However, this technique relies on ionizing radiation and comparatively expensive equipment, which makes it less feasible for repeated examinations to monitor liver disease progressions.

5.2.3.2. Magnetic resonance imaging

There are two basic types of MRI images called T_1-weighted and T_2-weighted images. Both mapping sequences can be used for tissue characterizations. T_1 mapping is a "promising diagnostic tool that has shown to be effective in differentiating different stages of fibrosis and also has shown potential for predicting clinical events. However, further research is needed to validate effective scoring systems and the influences of other compounding factors for it to become a valid alternative to liver biopsy in clinical practice" [212]. Additionally, MRI equipment is also very expensive to acquire and maintain.

5.2.3.3. Magnetic resonance elastography

Magnetic resonance elastography (MRE) can predict fibrosis stages "effectively in patients with chronic liver diseases, while producing a wider and more representative map of liver stiffness in both 2-D and 3-D planes" in comparison to other alternative methods [212]. MRE works similar to TE (see Section 5.2.2.2) by generating shear waves and imaging them with phase contrast MRI. MRE has been researched thoroughly in multiple studies but this technique still suffers from the main limitations of MRI, such as high acquisition and maintenance costs, as it has to rely on similar equipment (see Section 5.2.3.2).

5.2.3.4. Magnetic resonance spectroscopy

Magnetic resonance spectroscopy (MRS) "evaluates proton signals as a function of their resonance frequency, which shows multiple peaks at different locations within a specified volume of the liver. The MR spectrum describes the intensity of MR signals as a function of precession frequency, with fat and water producing the most visible peaks". A fatty liver presents with spectral peaks for water and fat, while a non-fatty liver shows only a peak due to the presence of water [215]. As described above for MRI and MRE, the equipment needed to perform MRS is also very expensive to acquire and maintain.

5.2.3.5. Magnetic resonance imaging proton density fat fraction

MRI-PDFF "measures the fraction of MRI-visible protons bound to fat divided by all MRI-visible protons in the liver (fat and water). Using this technique, the liver signal [obtained via MRI] is divided into water and fat signal components by acquiring gradient echoes at appropriately spaced echo times, so as to quantify the percentage of liver fat" [215]. Again, the equipment required to perform MRI-PDFF is expensive w.r.t to acquisition costs and maintenance.

5.2.4. Comparison of different methods

All non-invasive techniques not based on US have one disadvantage in common: The respective equipment needed to perform measurements is usually very complex and expensive to acquire and maintain. Hence, these techniques are unsuitable for mobile or wearable applications. Most devices leveraging US signals share this disadvantage. All solutions based on ARFI (see Section 5.2.2.1), TE (see Section 5.2.2.2) or SWE (see Section 5.2.2.3) currently require sophisticated equipment to generate and detect shear waves. 2-D B-Mode US with modern high-end devices (see Section 5.2.2.4) are a viable alternative to some extent. However, there are several drawbacks of conventional B-Mode US for NAFLD evaluations [215]:

1. “It is qualitative and therefore subjective” as there is a “lack of sonographic criteria for different degrees of steatosis”.
2. “Sensitivity is limited when there are few steatotic hepatocytes”.
3. “Sensitivity and specificity of B-Mode sonography decreases as body mass index increases”.
4. “Conventional sonography cannot differentiate [between] steatosis and steatohepatitis or stage fibrosis”.

QUS (see Section 5.2.2.5) might be a better choice to address those shortcomings but this technique still needs to be further investigated w.r.t. accuracy, reproducibility and repeatability. Hence, when it comes to low-cost and mobile solutions, the only viable option remaining is A-Mode US (see Section 5.2.2.6). Even though promising initial results for reliable predictions based on those signals exist [25], further research is needed. The following sections provide a thorough discussion of the suitability of this approach by describing the analysis of signals acquired from a clinical cohort.

5.3. Materials and Methods

Section 5.3.1 provides details of the clinical study the results in this chapter are based on. Section 5.3.2 illustrates the complete data acquisition process, while Section 5.3.3 provides an overview of how the signals were processed before using them as input data for the ML models mentioned in Section 5.3.5. Section 5.3.4 provides more information about the different deployed annotation schemes.

5.3.1. Clinical study

The 1-D US signals used for the classification of liver steatosis grades and liver fibrosis stages were obtained from patients participating in a clinical study conducted at the *J. W. Goethe University Hospital* in Frankfurt, Germany. The database consists of a cohort with signals from 27 patients examined between May 2019 and March 2020 by an experienced radiographer. The subjects all provided written consent, were between 26 and 70 years old and presented with different fibrosis and steatosis stages. Most patients suffered from NAFLD or NASH. A commercially available *Siemens Acuson S2000* device, which allows access to the raw 1-D US signals, was used to acquire all data. See Appendix C for a comprehensive overview of all participants of this clinical study.

In addition to raw 1-D US signals, experienced clinicians also obtained further data used for ground truth annotations. These included fibrosis and steatosis assessments based on *CAP* values acquired with a Fibroscan device or US speed values acquired with an *ARFI* device.

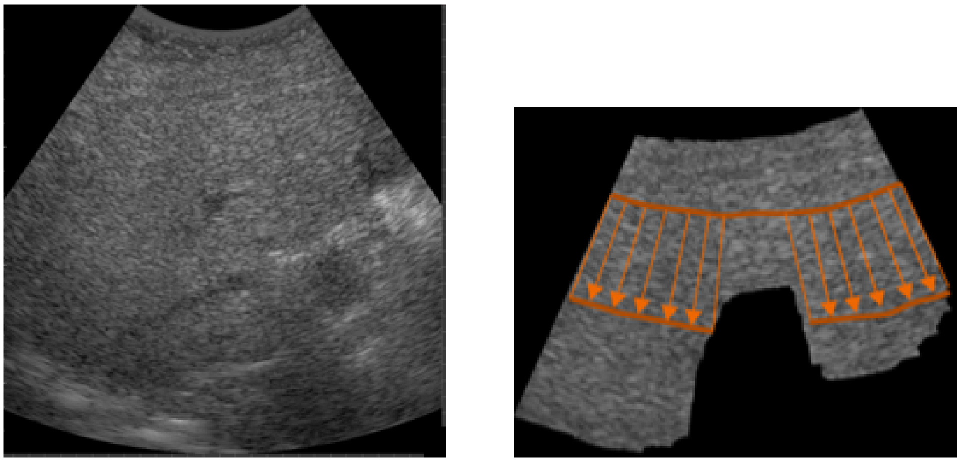
5.3.2. Data acquisition

The underlying 1-D US signals used for the ML models in this work are the result of manual segmentation of the input data. This manual segmentation was necessary to ensure that no signals from the surrounding vasculature in the liver parenchyma were present in the final input data as these signals would have impacted the output of the ML models designed to classify signals of interest stemming from the liver only. Figure 5.1 (a) depicts the B-Mode reconstruction of the initial input data while Figure 5.1 (b) shows what the segmented data looks like. The orange overlay illustrates the depth intervals from which the final A-scans where extracted. Note that the B-Mode reconstruction was only used to perform the segmentation and not as input data for the ML pipeline. Figure 5.2 shows an example of an extracted A-Scan, while Figure 5.3 depicts the truncated version of the same A-Scan that has been cropped to a signal depth of 3 cm to 6 cm as the liver is typically located within this depth in the body. The x-axis of both A-scans shows the signal duration in µs, while the y-axis shows the signal amplitude.

Subcostal means “below a rib or the ribs” while *intercostal* means “between the ribs”. The cohort contains signals acquired via a subcostal position from the left liver lobe, signals acquired via a subcostal position from the right liver lobe and signals acquired via an intercostal position from the right liver lobe. The time gain compensation, which accounts for tissue attenuation, was either set to “diagnostic” or “maximum constant”. The former was the most fitting setting to obtain a result with comparatively homogeneous amplitude values, which is most often used for diagnostic imaging. The signals were acquired by following a protocol including two iterations. In total, 334,721 A-Scans were acquired. Table 5.3 provides a summary of the total amounts of A-Scans for different liver lobes and US transducer positions.

5.3.3. Data preprocessing

To analyze and classify the obtained 1-D US signals, they can be preprocessed with a variety of different approaches. For example, statistical, temporal and spectral feature extraction approaches (see Section 2.3.3) or



(a) 2-D reconstruction of 1-D input datafrom the liver.

(b) Manually segmented input data. The orange overlay illustrates the depth intervals from which the final A-Scans were extracted.

Figure 5.1.: 2-D reconstruction of 1-D input data (left) and manually segmented input data of the liver (right).

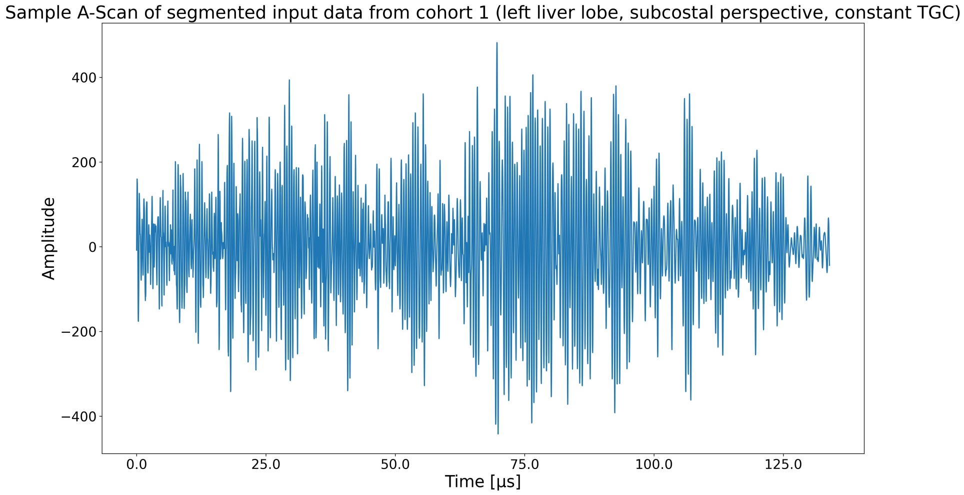


Figure 5.2.: Example of an extracted A-Scan.

signal transformations (see Section 2.3.2) might be used for signal pre-processing. For the analyses presented in this chapter, possible input data types were:

1. Raw 1-D US signals
2. 1-D US spectral representation computed using a Fast Fourier transform (see Section 2.3.2.1)
3. 1-D US signals transformed with a Wavelet transform (see Section 2.3.2.2)

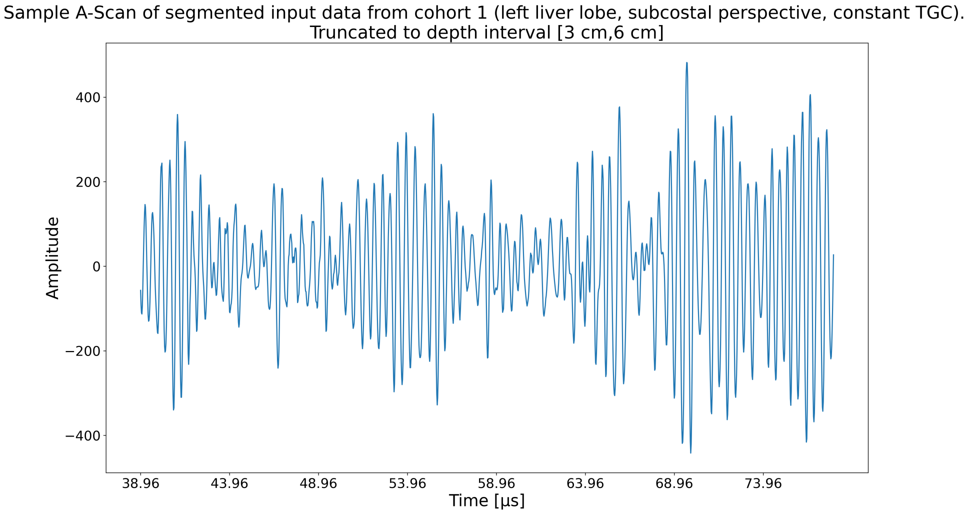


Figure 5.3.: Example of an extracted A-Scan truncated to a depth interval of 3 cm to 6 cm.

Liver lobe	Position	Total A-Scans	Total A-Scans (with segmentation in relevant depth interval)
left	subcostal	109,809	48,206
right	subcostal	112,113	80,716
right	intercostal	112,799	81,366

Table 5.3.: Summary of total A-Scans for different liver lobes and transducer positions.

4. 1-D US signals transformed with a Hilbert transform (see Section 2.3.2.3)
5. Band-pass filtered 1-D US signals (see Section 2.3.2.4)
6. Statistical features extracted from raw 1-D US signals (see Section 2.3.3.1)
7. Temporal features extracted from raw 1-D US signals (see Section 2.3.3.2)
8. Spectral features extracted from raw 1-D US signals (see Section 2.3.3.3)
9. A combination of statistical, temporal and spectral features extracted from raw 1-D US signals

5.3.4. Annotation schemes

This work explores different annotation schemes depending on the classification one would like to achieve. All annotation schemes are based on additional data acquired during the clinical study described in Section 5.3.1. For liver steatosis assessments, the annotation scheme relied on *CAP* values (in db/m) acquired with the TE method used in the Fibroscan device. This chapter describes the analysis of three possible annotation schemes for the steatosis stages by grouping the signals for binary annotations: either “S0 vs. S1, S2, S3”, “S0, S1 vs. S2, S3” or ”S0, S1, S2 vs. S3”. The fibrosis annotation schemes either relied on CAP values or US speed values (in m/s) acquired with an *ARFI* device. To this end, two possible annotation schemes for fibrosis assessments either grouped the signals for binary annotations with the scheme ”<F2 vs. F2, F3, F4” or “≤F2 vs. F3,F4”.

5.3.5. Data analysis

The ML approaches *LightGBM*, *XGBoost*, *LR*, *SVM* and *Transformers* classified the acquired signals. Alternatives, such as *1-NN DTW* or other *ANNs* were not used because of the very long run-times of those algorithms. Additionally, previous results presented in Chapter 3 and Chapter 4 did not provide any proof that those algorithms would have yielded substantially better classification performances.

Unfortunately, *FibroScan* values for two patients were missing (see Table C.4). To address this issue, a value imputation strategy provided by the *scikit-learn* library [124] has been implemented.

5.4. Results

This section presents results of DRTs, such as PCA and t-SNE (see Section 5.4.1) and results of a variety of ML models (see Section 5.4.2). The results of the latter are divided into ML models classifying different fibrosis (see Section 5.4.2.1) and steatosis (see Section 5.4.2.2) stages. Please note that those results are again subdivided by results stemming from ML models based on 1-D US signals of the right and left liver lobes.

5.4.1. Dimensionality Reduction Techniques

5.4.1.1. PCA

This section presents PCA visualizations of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a diagnostic TGC and have been annotated using ARFI values (see Figure 5.4) or *FibroScan* values (see Figure 5.5) for the Fibrosis staging scheme ≤ F2 vs. F3, F4.

It is apparent that none of these PCA visualizations features a strict separation of the underlying signals.

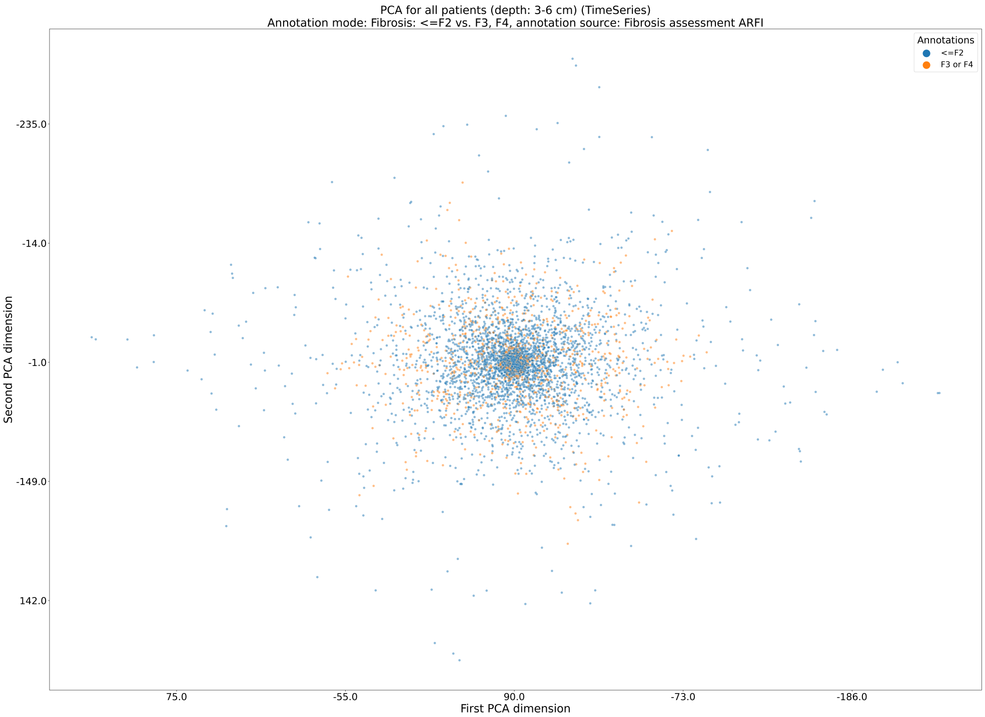


Figure 5.4.: PCA visualization of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a diagnostic TGC and have been annotated using ARFI values and the Fibrosis staging scheme ≤ F2 vs. F3, F4.

5.4.1.2. t-SNE

This section presents t-SNE visualizations of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a diagnostic (see Figure 5.6) or constant (see Figure 5.7) TGC and have been annotated using ARFI values or *FibroScan* values (see Figure 5.8 and Figure 5.9) for the Fibrosis staging scheme ≤ F2 vs. F3, F4.

It is apparent that none of these selected t-SNE visualizations features a strict separation of the underlying signals. However, Figure 5.8 shows a general tendency of signals grouping according to their respective annotations.

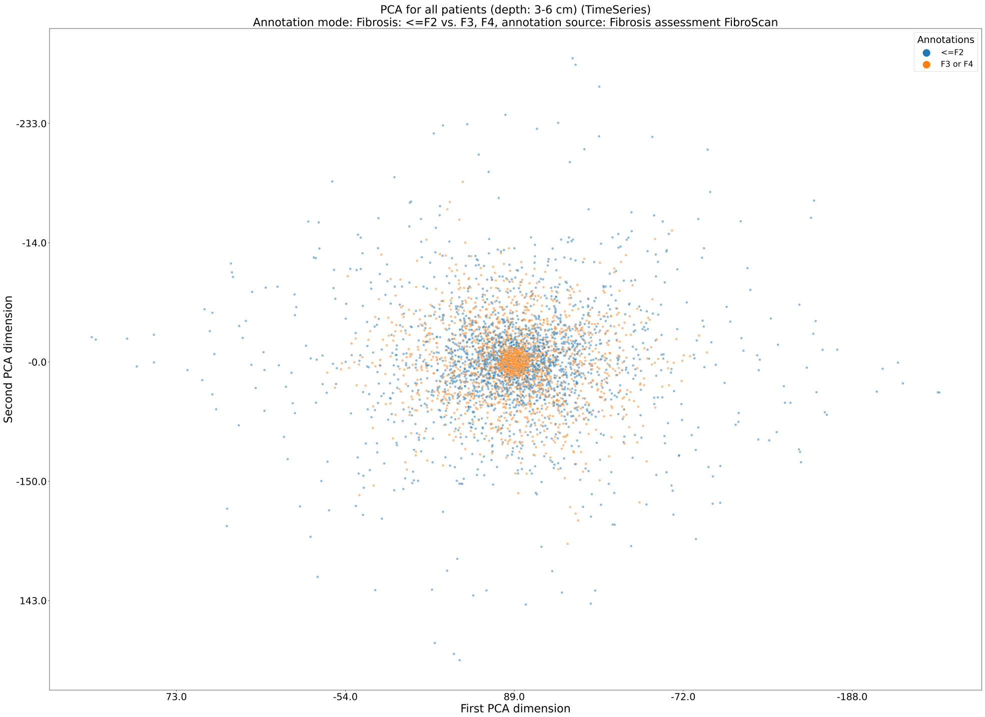


Figure 5.5.: PCA visualization of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a diagnostic TGC and have been annotated using *FibroScan* values and the Fibrosis staging scheme ≤ F2 vs. F3, F4.

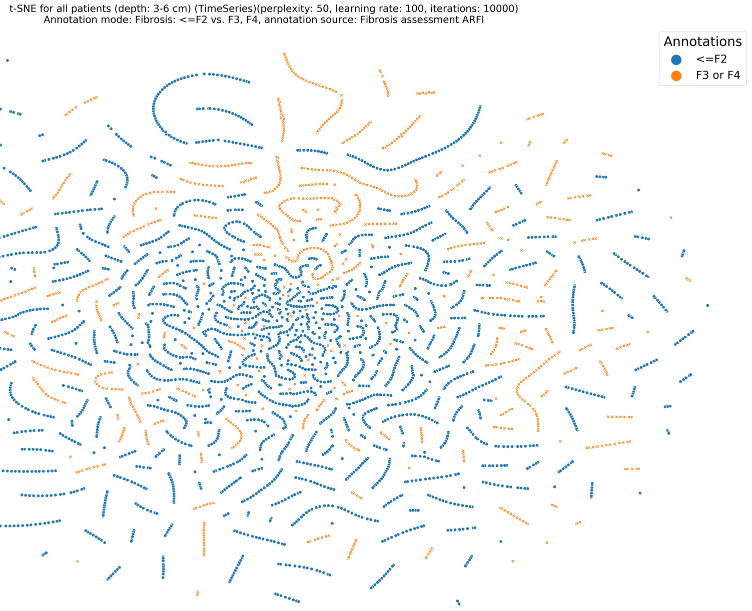


Figure 5.6.: t-SNE visualization of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a diagnostic TGC and have been annotated using ARFI values and the Fibrosis staging scheme ≤ F2 vs. F3, F4.

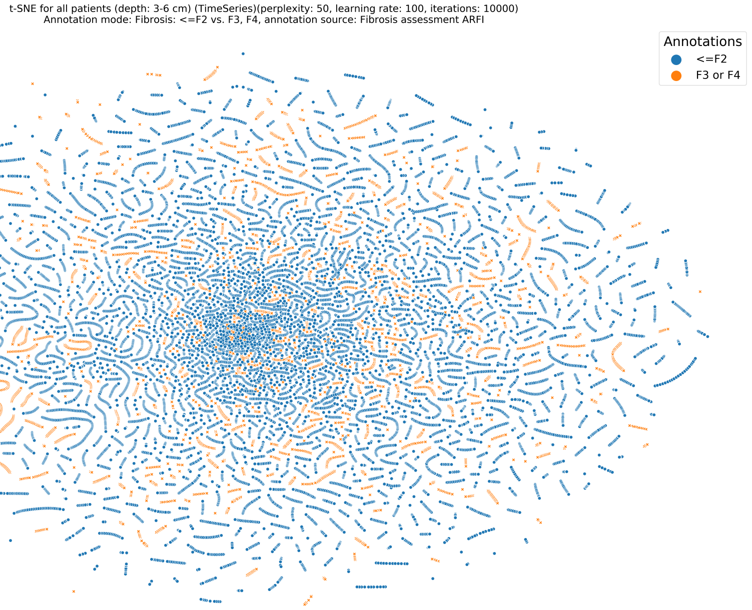


Figure 5.7.: t-SNE visualization of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a constant TGC and have been annotated using ARFI values and the Fibrosis staging scheme ≤ F2 vs. F3, F4.

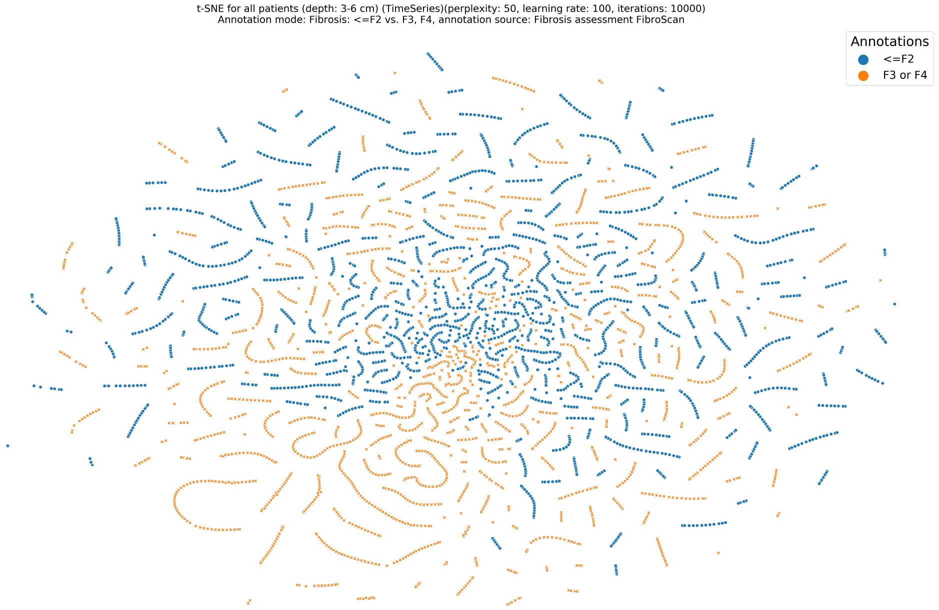


Figure 5.8.: t-SNE visualization of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a diagnostic TGC and have been annotated using *FibroScan* values and the Fibrosis staging scheme ≤ F2 vs. F3, F4.

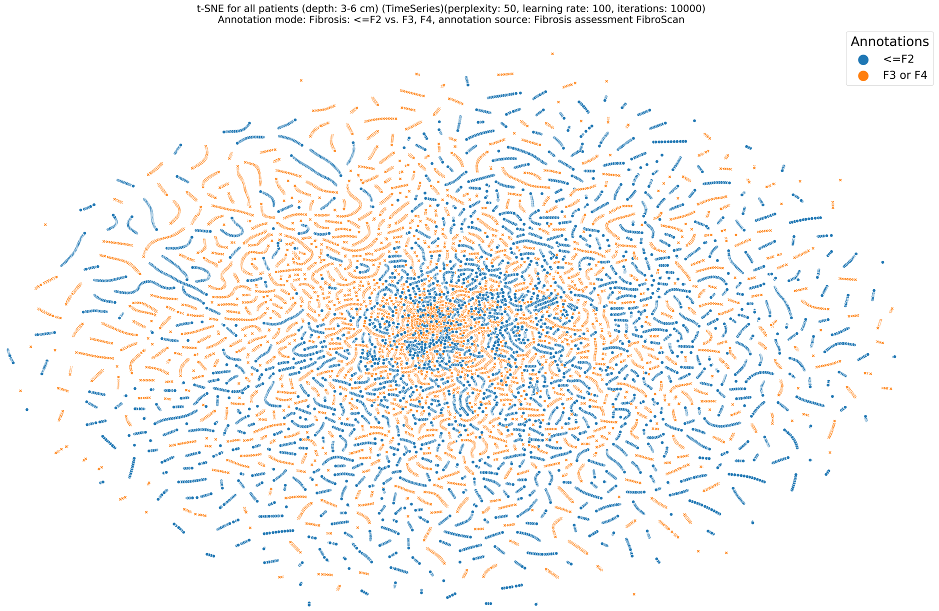


Figure 5.9.: t-SNE visualization of 1-D US signals stemming from the depth interval [3,6] cm of the left liver lobe acquired from a subcostal perspective. The signals were acquired from the first protocol iteration with a constant TGC and have been annotated using *FibroScan* values and the Fibrosis staging scheme ≤ F2 vs. F3, F4.

5.4.2. Machine Learning

5.4.2.1. Fibrosis

Signals stemming from left liver lobe

The following tables include results of ML models based on 1-D US signals stemming from the left liver lobe only.

≤ F2 vs. F3, F4 annotation scheme

Table 5.4 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the fibrosis stage ≤ *F2 vs. F3, F4* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 85.71% has been achieved with a SVM model based on Wavelet transformed 1-D US signals acquired from the second iteration of acquisitions taken from a subcostal US transducer perspective with a diagnostic TGC stemming from the left liver lobe. The respective annotations were created with categories based on ARFI values.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F_1-score (%)
SVM	Wavelet transformed A-scans	ARFI	2	subcostal	diagnostic	9.31	85.71 %
SVM	Statistical TSFEL features	ARFI	2	subcostal	diagnostic	32.50	80.94 %
SVM	Raw A-scans	ARFI	2	subcostal	diagnostic	10.57	78.52 %
SVM	Bandpassed A-scans	ARFI	2	subcostal	diagnostic	11.66	73.78 %
LightGBM	Raw A-scans	ARFI	1	subcostal	constant	482.45	63.56 %

Table 5.4.: ≤ F2 vs. F3, F4 annotation scheme results for signals stemming from left liver lobe

< F2 vs. F2, F3, F4 annotation scheme

Table 5.5 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the fibrosis stage < *F2 vs. F2, F3, F4* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 52.12% has been achieved with a SVM model based on statistical TSFEL features extracted from 1-D US signals acquired from the second iteration of acquisitions taken from a subcostal US transducer perspective with a diagnostic TGC stemming from the left liver lobe. The respective annotations were created with categories based on ARFI values.

Signals stemming from right liver lobe

The following tables include results of ML models based on 1-D US signals

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F_1-score (%)
SVM	Statistical TSFEL features	ARFI	2	subcostal	diagnostic	32.64	52.12
LightGBM	Spectral TSFEL features	FibroScan	1	subcostal	diagnostic	244.56	42.66
SVM	Statistical TSFEL features	FibroScan	1	subcostal	constant	216.92	42.00
LightGBM	Temporal TSFEL features	FibroScan	1	subcostal	diagnostic	50.07	41.55
XGBoost	Fourier transformed A-scans	FibroScan	1	subcostal	diagnostic	11.83	41.48

Table 5.5.: < F2 vs. F2, F3, F4 annotation scheme results for signals stemming from left liver lobe

stemming from the right liver lobe only.

≤ F2 vs. F3, F4 annotation scheme

Table 5.6 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the fibrosis stage ≤ *F2 vs. F3, F4* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 80.77% has been achieved with a Transformer model relying on a combination of all extracted TSFEL features from 1-D US signals acquired from the first iteration of acquisitions taken from an intercostal US transducer perspective with a diagnostic TGC stemming from the right liver lobe. The respective annotations were created with categories based on ARFI values.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F1-score
Transformer	All TSFEL features combined	ARFI	1	intercostal	diagnostic	799.46	80.77
Transformer	Statistical TSFEL features	ARFI	1	intercostal	diagnostic	459.47	80.77
Transformer	All TSFEL features combined	ARFI	1	subcostal	diagnostic	926.98	77.78
Transformer	Statistical TSFEL features	ARFI	1	subcostal	diagnostic	537.26	77.78
SVM	Temporal TSFEL features	ARFI	1	subcostal	diagnostic	100.78	75.89

Table 5.6.: ≤ F2 vs. F3, F4 annotation scheme results for signals stemming from right liver lobe

< F2 vs. F2, F3, F4 annotation scheme

Table 5.7 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the fibrosis stage < *F2 vs. F2, F3, F4* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 69.23% has been achieved with a Transformer model based on a combination of all TSFEL features extracted from 1-D US signals acquired from the first iteration of acquisitions taken from an intercostal US transducer perspective with a diagnostic TGC stemming from the right liver lobe. The respective annotations were created with categories based on ARFI values.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F1-score
Transformer	All TSFEL features combined	ARFI	1	intercostal	diagnostic	798.70	69.23
Transformer	Spectral TSFEL features	ARFI	1	intercostal	diagnostic	820.94	69.23
Transformer	All TSFEL features combined	ARFI	1	subcostal	diagnostic	918.35	66.67
Transformer	Statistical TSFEL features	ARFI	1	subcostal	diagnostic	537.82	66.67
SVM	Statistical TSFEL features	FibroScan	1	intercostal	constant	749.28	51.79

Table 5.7.: < F2 vs. F2, F3, F4 annotation scheme results for signals stemming from right liver lobe

5.4.2.2. Steatosis

Signals stemming from left liver lobe

The following tables include results of ML models based on 1-D US signals stemming from the left liver lobe only.

S0 vs. S1, S2, S3 annotation scheme

Table 5.8 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the steatosis stage *S0 vs. S1, S2, S3* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 80.95% has been achieved with a LightGBM model based on Bandpassed 1-D US signals acquired from the second iteration of acquisitions taken from a subcostal US transducer perspective with a diagnostic TGC stemming from the left liver lobe. The respective annotations were created with categories based on CAP values acquired with a FibroScan system.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F_1-score (%)
LightGBM	Bandpassed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	261.32	80.95
LightGBM	Hilbert transformed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	235.74	80.95
LightGBM	Raw A-scans	CAP (FibroScan)	2	subcostal	diagnostic	218.98	80.95
LightGBM	Wavelet transformed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	102.33	80.95
SVM	Bandpassed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	12.14	80.95

Table 5.8.: S0 vs. S1, S2, S3 annotation scheme results for signals stemming from left liver lobe

S0, S1 vs. S2, S3 annotation scheme

Table 5.9 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the steatosis stage *S0, S1 vs. S2, S3* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 80.95% has been achieved with a LightGBM model based on Bandpassed 1-D US signals acquired from the second iteration of acquisitions taken from a subcostal US transducer perspective with a diagnostic TGC stemming from the left liver lobe. The respective annotations were created with categories based on CAP values acquired with a FibroScan system.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F_1-score (%)
LightGBM	Bandpassed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	261.11	80.95
LightGBM	Hilbert transformed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	235.87	80.95
LightGBM	Raw A-scans	CAP (FibroScan)	2	subcostal	diagnostic	219.10	80.95
LightGBM	Wavelet transformed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	102.29	80.95
SVM	Bandpassed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	12.12	80.95

Table 5.9.: S0, S1 vs. S2, S3 annotation scheme results for signals stemming from left liver lobe

S0, S1, S2 vs. S3 annotation scheme

Table 5.10 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the steatosis stage *S0, S1, S2 vs. S3* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 76.19% has been achieved with a LightGBM model based on Bandpassed 1-D US signals acquired from the second iteration of acquisitions taken from a subcostal US transducer perspective with a diagnostic TGC stemming from the left liver lobe. The respective annotations were created with categories based on CAP values acquired with a FibroScan system.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F1-score
LightGBM	Bandpassed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	265.29	76.19
LightGBM	Raw A-scans	CAP (FibroScan)	2	subcostal	diagnostic	218.31	76.19
LightGBM	Wavelet transformed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	102.01	76.19
SVM	Bandpassed A-scans	CAP (FibroScan)	2	subcostal	diagnostic	12.36	76.19
SVM	Statistical TSFEL features	CAP (FibroScan)	2	subcostal	diagnostic	32.66	76.19

Table 5.10.: S0, S1, S2 vs. S3 annotation scheme results for signals stemming from left liver lobe

Signals stemming from right liver lobe

The following tables include results of ML models based on 1-D US signals stemming from the right liver lobe only.

S0 vs. S1, S2, S3 annotation scheme

Table 5.11 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the steatosis stage *S0 vs. S1, S2, S3* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 81.48% has been achieved with a SVM model based on statistical TSFEL features extracted from 1-D US signals acquired from the first iteration of acquisitions taken from an intercostal US transducer perspective with a constant TGC stemming from the right liver lobe. The respective annotations were created with categories based on CAP values acquired with a FibroScan system.

S0, S1 vs. S2, S3 annotation scheme

Table 5.12 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the steatosis stage *S0, S1 vs. S2, S3* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 77.78% has been achieved with a SVM model based on statistical TSFEL features extracted from 1-D US signals acquired from the first iteration of acquisitions taken from an intercostal US transducer perspective with a constant TGC stemming from the right liver lobe. The respective annotations were created with categories based on CAP values acquired with a FibroScan system.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F1-score
SVM	Statistical TSFEL features	CAP (FibroScan)	1	intercostal	constant	744.61	81.48
SVM	Statistical TSFEL features	CAP (FibroScan)	2	intercostal	diagnostic	102.67	81.48
SVM	Raw A-scans	CAP (FibroScan)	2	intercostal	diagnostic	53.84	81.48
SVM	Wavelet transformed A-scans	CAP (FibroScan)	2	intercostal	diagnostic	39.37	81.48
SVM	Statistical TSFEL features	CAP (FibroScan)	2	intercostal	constant	892.34	81.48

Table 5.11.: S0 vs. S1, S2, S3 annotation scheme results for signals stemming from right liver lobe

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F1-score
SVM	Statistical TSFEL features	CAP (FibroScan)	1	intercostal	constant	750.12	77.78
SVM	Statistical TSFEL features	CAP (FibroScan)	2	intercostal	diagnostic	102.83	77.78
SVM	Statistical TSFEL features	CAP (FibroScan)	2	intercostal	constant	899.12	77.78
SVM	Statistical TSFEL features	CAP (FibroScan)	1	subcostal	diagnostic	102.25	77.78
SVM	Statistical TSFEL features	CAP (FibroScan)	1	intercostal	diagnostic	90.07	76.92

Table 5.12.: S0, S1 vs. S2, S3 annotation scheme results for signals stemming from right liver lobe

S0, S1, S2 vs. S3 annotation scheme

Table 5.13 presents the five best performing combinations of ML model, data type, assessment method, protocol iteration, US transducer position and TGC mode for the steatosis stage *S0, S1, S2 vs. S3* annotation scheme and lists the time needed for training and evaluation and the resulting average F_1-score for each combination. The best average F_1-score of 74.07% has been achieved with a SVM model based on statistical TSFEL features extracted from 1-D US signals acquired from the first iteration of acquisitions taken from an intercostal US transducer perspective with a constant TGC stemming from the right liver lobe. The respective annotations were created with categories based on CAP values acquired with a FibroScan system.

ML model	Data type	Assessment method	Protocol iteration	US transducer perspective	Time gain compensation	Time for training and evaluation (in minutes)	Average F1-score
SVM	Statistical TSFEL features	CAP (FibroScan)	1	intercostal	constant	771.72	74.07
SVM	Statistical TSFEL features	CAP (FibroScan)	2	intercostal	constant	925.22	74.07
SVM	Statistical TSFEL features	CAP (FibroScan)	1	subcostal	diagnostic	102.76	74.07
SVM	Statistical TSFEL features	CAP (FibroScan)	1	intercostal	diagnostic	90.80	73.08
SVM	Statistical TSFEL features	CAP (FibroScan)	2	intercostal	diagnostic	103.46	72.22

Table 5.13.: S0, S1, S2 vs. S3 annotation scheme results for signals stemming from right liver lobe

5.5. Discussion of liver steatosis and liver fibrosis classifications

Neither of the DRTs PCA or t-SNE provides a clear visual separation of the examined 1-D US signals. Nevertheless, the following sections describe classifications of those signals for fibrosis (see Section 5.5.1) and steatosis (see Section 5.5.2) staging.

5.5.1. Fibrosis stage classifications

For fibrosis stage classifications, the best average F_1-score of 85.71% has been achieved with a SVM model based on Wavelet transformed 1-D US signals acquired from the second iteration of acquisitions taken from a subcostal US transducer perspective with a diagnostic TGC stemming from the left liver lobe. The respective annotations were created with categories based on ARFI values and the annotation scheme ≤ F2 vs. F3, F4 (see Table 5.4). The training and evaluation of this combination took 9.31 minutes to complete. The performances of all other ML model / data type combinations and all other annotations were significantly worse. Note that a SVM model even outperformed more recent algorithms, such as LightGBM or Transformers.

5.5.2. Steatosis stage classifications

For steatosis stage classifications, the best average F_1-score of 81.48% has been achieved with a SVM model. This model either used statistical TSFEL features, Wavelet transformed 1-D US signals or raw 1-D US signals from various protocol iterations. The signals were acquired from an intercostal US transducer perspective with a diagnostic or constant TGC (see Table 5.11). The respective annotation scheme S0 vs. S1, S2, S3 was based on CAP values acquired with the *FibroScan* device. The training and evaluation of the fastest combination took 39.37 minutes to complete. The performances of all other ML model / data type combinations and all other annotations were worse. Note that a SVM model even outperformed more recent algorithms, such as LightGBM or Transformers.

5.5.3. General discussion

In this work, fibrosis and steatosis stages can both be classified with a high average F_1-score. F_1 scores > 80% for steatosis (distinguishing between the classes S0 and S1, S2, S3) and fibrosis (distinguishing between the classes ≤ F2 and F3, F4) have been achieved in less than an hour on signals stemming from 27 subjects. Remarkably, these results have been achieved with the SVM algorithm, which has been used for decades. More sophisticated ML models, such as a Transformer model or GBMs, do not yield this performance.

5.5.3.1. Limitations

Even though this work presents promising results, certain limitations remain. The presented results can only be achieved after manually pre-processing the input signals as described in Section 5.3. This remains an obstacle to future deployments of the presented algorithms in commercial devices. Additionally, the clinical study presented in this work was comparatively small. It is highly likely that this has contributed to results that are not yet suitable for commercial devices in clinical everyday use, even though they are very promising.

5.5.3.2. Future works

It has been shown that “in recent years, significant progress has been made in developing more accurate and efficient machine learning algorithms for segmentation of medical and natural images” [218]. Hence, future research should include automatic image segmentation of the input data to segment between liver parenchyma and the surrounding soft tissue to ensure that no inference from the latter influences the signals.

Additionally, a larger clinical study is needed to validate the results of this work and to achieve average F_1-scores that would justify a potential certification by medical authorities, such as the FDA, for devices suitable for everyday clinical use.

Chapter 6

Discussion

The major goal of this work is to present an automatic and domain agnostic approach to classify 1-D US signals, which requires as little domain specific knowledge (DSK) as possible for the end-user. The results presented in Section 3.4, Section 4.4 and Section 5.4 indicate that effective classifications of 1-D US signals to distinguish between different *muscle contraction states*, *muscle fatigue states*, *epiphyseal growth plate closure states* and *liver disease states* are possible with sufficient degrees of accuracy. The following sections discuss these results.

Contents

6.1. Comparison

Before discussing classifications of 1-D US signals based on ML for a variety of different tasks, this section first compares 1-D A-Mode US to medical imaging techniques and presents significant advantages and disadvantages. The reliance on signals acquired with non-ionizing radiation has several significant advantages over alternative methods that might be used for the tasks described in this work. The low-cost hardware needed to acquire these signals and the possibility of obtaining information from deeper soft tissue layers are additional advantages of this approach. Table 6.1 presents a compact comparison of different methods usable in the medical field.

Method	Harm potential	Low-cost	Wearability	Intuitiveness of underlying data for humans	Availability of information from deep tissue layers	Applicable for muscle state classifications	Applicable for epiphyseal radius bone closure detection	Applicable for liver disease state classifications
1-D ultrasound (A-Mode)	+	+	+	0	+	+	+	+
2-D ultrasound (B-Mode)	+	+	+	+	+	+	+	+
Ultrasound elastography	+	0	-	0	+	+	-	+
Medical X-ray imaging	0	0	-	+	+	-	+	+
Magnetic resonance imaging	+	-	-	+	+	+	+	+
Electromyography (EMG)	+	+	+	0	-	+	-	-
Surface Electromyography (sEMG)	+	+	+	0	-	+	-	-
Mechanomyography (MMG)	+	+	+	0	-	+	-	-
Textile resistive pressure mapping sensors	+	+	+	0	-	+	-	-
Inertial measurements units (IMUs)	+	+	+	0	-	+	-	-

Table 6.1.: Comparison of different methods usable for the tasks presented in this work. The colors indicate the suitability for the property described in the column heading. Green indicates a general suitability, while gray states that significant disadvantages exist. A red cell means that a method is not suitable for a given scenario.

1-D US signals contain information from deeper soft tissue layers and can be acquired non-invasively with low-cost equipment, which makes mobile or wearable devices possible. Significant disadvantages are that the acquisition of 1-D US signals requires certain preparation steps (see Section 6.4.1) and that the signals are not intuitively interpretable for humans without the help of sophisticated DSP or ML algorithms.

6.2. Interpretation

This work shows that 1-D US signal classification in general is possible with widely available TSC algorithms. In contrast to more traditional DSP methods, stand-alone ML methods or approaches combining DSP and ML are preferable as real-time scenarios require very fast inference times that can be achieved using pre-trained models. To this end, this work focuses on solutions exploiting ML. It enriches the field of 1-D US signal classifications by examining three major fields in detail.

A variety of publications investigate the classification of medical images in general or 2-D US B-Mode images in particular but 1-D US signal classification approaches have, so far, often been overlooked.

The results achieved in this work prove that it is possible to classify 1-D US signals with high accuracies, even if those signals do not stem from carefully selected sources of interest, such as very pronounced liver areas or pre-selected muscle groups. Muscle activity 1-D US signals have been acquired in other works with multiple single-element US transducers, a system consisting of a receiver and transmitter part or with a single element US transducer in combination with a prior 2-D B-Mode investigation finding the best possible transducer position for the examined muscles (see Section 3.2.7). In contrast to that, this work only relies on signals stemming from a single element US transducer without conducting prior B-Mode investigations to obtain high-quality signals. This approach facilitates low-cost solutions requiring very little electric energy. By incorporating DSK to extract features or prepare the signals, respectively, the complexity of ML models can be reduced. Low-cost hardware and ML models with a low complexity allow mobile and wearable scenarios with cloud-based or Edge AI applications. The former uses the local device to send signals to a cloud computer, responsible for training the respective ML models and / or incorporating newly available signals into existing pipelines. The latter performs all necessary steps, such as training or evaluating the ML models, on the local device, which has the advantages of low latencies, high data security, more independence from remote infrastructure and more efficient use of network bandwidth.

The approaches presented in this work serve as a strong foundation for future research to enrich or substitute expensive, stationary or invasive solutions in a variety of different domains, such as medicine, sports or rehabilitation. Chapter 3 shows the potential 1-D US signal classifications have for the mobile and wearable recognition of muscle contraction and fatigue states. A device exploiting this approach could provide an alternative to existing solutions that manually track the current state of a subject's fitness or rehabilitation progress. Chapter 4 provides insights on how the presented classification models can substitute potentially harmful and ionizing examinations based on X-rays for epiphyseal growth plate closure detection, while Chapter 5 describes the possibility of equipping laypersons with a system incorporating medical expert knowledge to perform liver state classifications with a low-cost smart mobile device based on ML models using 1-D US signals.

6.3. Implications

Table 6.2 summarizes the best performing ML model / data type combinations for different scenarios of this work. Results for muscle state classifications, epiphyseal radius bone closure detections and liver disease stage classifications are depicted in red, green and yellow, respectively.

Section 6.3.1 and Section 6.3.2 discuss the suitability of different ML models and data types.

Scenario	Model	Data type	Time for training and evaluation	Average F_1-score (%)
Muscle contraction classifications	SVM	Hilbert transformed signals	~10 minutes	88
Muscle fatigue classifications on signals stemming from the dominant arm of female subjects only	Logistic Regression	Extracted spectral features	<5 minutes	86
Muscle fatigue classifications on signals stemming from the dominant arm of male subjects only	SVM	Wavelet transformed signals	5 minutes	86
Epiphyseal radius bone closure detection	CatBoost	Raw ultrasound signals	26.08 hours	87
Epiphyseal radius bone closure detection	XGBoost	Raw ultrasound signals	<5 minutes	85
Liver fibrosis stage detection on signals stemming from the left liver lobe (subcostal perspective)	SVM	Wavelet transformed signals	<10 minutes	86
Liver steatosis stage detection on signals stemming from the rightliver lobe (intercostal perspective)	SVM	Wavelet transformed signals	~40 minutes	81

Table 6.2.: Summary of best performing ML model / data type combinations for different scenarios of this work. Results for muscle state classifications, epiphyseal radius bone closure detections and liver disease stage classifications are depicted in red, green and yellow, respectively.

6.3.1. Suitability of different ML models

Previous works by other researchers have found high accuracies in TSC tasks using classic 1-NN DTW [80, 81]. However, this work finds the classic 1-NN DTW algorithm is not the leading model in terms of speed or accuracy in any scenario in which it was deployed (see Section 3.4 and Section 4.4) and should no longer be considered a benchmark algorithm for TSC tasks based on 1-D US signals.

Another observation is that the algorithms LR and SVM perform extraordinarily well in comparison to other, much more complex and recent, methods in certain scenarios. The SVM algorithm, in its current form, has been published first in 1964 [144], while the roots of LR reach back even further. Despite having been deployed for many decades, these methods still outperform all other evaluated models for a variety of tested data types for muscle state classifications (see Section 3.4). In some cases, LR and SVM yield top performances for the identification of hepatic steatosis and fibrosis in patients with NAFLD (see Section 5.4). However, this is not always the case in this scenario as Transformers or GBMs also outperform these algorithms in some cases. When it comes to the detection of epiphyseal radius bone closure, neither LR nor SVM yield top performances (see Section 4.4). These results emphasize that classic ML algorithms should always be considered for the classification of 1-D US signals.

It is notable that GBMs do not appear in the best five performing models for each muscle state classifications scenario (see Section 3.4). However, these models perform extraordinary well for the detection of epiphyseal radius bone closure (see Section 4.4) and, in some cases, yield competitive results for the identification of hepatic steatosis and fibrosis in patients with NAFLD (see Section 5.4).

More traditional and comparatively recent 1-D ANN algorithms, such as MLP, FCN, ResNet or models of the ROCKET family do not perform well for muscle state classifications (see Section 3.4) or for the detection of

epiphyseal radius bone closure (see Section 4.4). Only the ANN Transformer method can, in some cases, achieve superior accuracies for the identification of hepatic steatosis and fibrosis in patients with NAFLD (see Section 5.4).
Hence, it is important to emphasize that a single ML model outperforming all other ML models for 1-D US signal classifications, regardless of the underlying application, does not exist. In 2021, the renowned company *DeepMind* published work describing a model that builds upon Transformers, which is "competitive with or outperforms strong, specialized models on classification tasks across various modalities: images, point clouds, audio, video, and video+audio". This model "scales to more than a hundred thousand inputs [and] opens new avenues for general perception architectures that make few assumptions about their inputs and that can handle arbitrary sensor configurations, while enabling fusion of information at all levels" [219]. However, groundbreaking work on such an all-encompassing model being able to classify all kinds of signals has not been presented yet and, at the time of writing, remains a topic of active research.

6.3.2. Suitability of different data types

There is no single data type yielding best performances in all examined scenarios. However, most best performing combinations w.r.t accuracy and speed rely on signals that have been transformed before being used as input data (see Section 3.4, Section 4.4 and Section 5.4). Signal transformations or feature extraction mechanisms are important to reduce the dimensionality of the input signals or to make them more suitable for further processing.
For instance, the Wavelet transform or the Hilbert transform can either smooth the input signals by removing amplitude fluctuations or by discarding negative amplitude values (see Figure 3.8). These properties make those signals a feasible choice for many ML models and commonly yield good results. Extracted features are especially suitable for LR or GBMs and can yield competitive results (see Table 6.2).

6.3.3. Ecological considerations

In 2018, a publication showed that the computational power required for "various large AI training models had been doubling every 3.4 months since 2012". This trend of increasing demands on sophisticated hardware and electrical energy has been called "red AI". An increase in the following three factors have been found to contribute to red AI models: "the cost of executing the model on a single example; the size of the training dataset, which controls the number of times the model is executed; and the number of hyperparameter experiments, which controls how many times the model is trained" [220]. In contrast to red AI, the concept of "green AI" has been proposed, which "yields novel results without increasing computational cost, and ideally reducing it" [220]. Hence, the faster models of this work, yielding results in less time, can be considered to belong to the field of green AI in contrast to slowly training models belonging to the field of red AI. Table 6.3

compares the ML models of this work w.r.t. their estimated ecological impact.

Model	Estimated ecological impact
SVM	++
Logistic Regression	++
XGBoost	++
LightGBM	++
MINIROCKET	++
MultiRocket (with MINIROCKET kernels)	++
Radial Basis Function Neural Network	++
MultiRocket (with ROCKET kernels)	+
CatBoost	**0**
ROCKET	**-**
MLP	–
FCN	–
ResNet	–
Transformer	–
1-NN DTW	–

Table 6.3.: Comparison of ML models w.r.t. their estimated ecological impact

Traditional approaches, such as SVM or LR, do not only have a minimal ecological impact, but also show superior accuracies in many cases as described above. Hence, green AI yields results superior to red AI approaches in this work, which makes the presented approaches more sustainable.

6.4. Limitations

Even though this work vividly demonstrates the usefulness and potential of 1-D US signal classifications for major applications, certain limitations remain. The following sections provide an overview of the remaining limitations of this work. Section 6.4.1 summarizes limitations regarding device technology, while Section 6.4.2 discusses limits regarding the used ML models.

6.4.1. Technical considerations for wearables

The wearable acquisition of 1-D US signals requires a rather complex handling, as the reflection of incident US waves at a surface boundary is proportional to the difference in acoustic impedance between both media. A coupling medium displaces air at the boundary surface of transducer and soft tissue and has an acoustic impedance much closer to the latter. This acoustic impedance matching greatly increases the intensity of the transmitted waves. Recent work has shown that “a conformal skin-worn device capable of parallel monitoring of [blood pressure], [heart rate] and multiple biomarkers” is possible by “addressing major engineering challenges in the integration of rigid ultrasound transducers and soft and stretchable electrochemical sensors into a single flexible and stretchable device while ensuring mechanical performance and avoiding signal crosstalk" [221]. More recently, the prototype of a "skin-conformal ultrasonic phased array for the monitoring of haemodynamic signals from tissues up to 14 cm beneath the skin" has been reported. "The device allows for active focusing and steering of ultrasound beams over a range of incident angles so as to target regions of interest" and "can be used to monitor Doppler spectra from cardiac tissues, record central blood flow waveforms and estimate cerebral blood supply in real time". However, this prototype "still requires wired data outputs to the back-end acquisition system for post-processing". Additional challenges, such as "on-board signal pre-conditioning, memory and wireless data transmitting, and replacing the power supply with a state-of-the-art flexible lithium-polymer battery" also still remain to be addressed in the future [59]. Apart from those challenges, this prototype allows a seamless integration removing "air gaps at the device-skin interface, which eliminates the requirement for ultrasound gels typically used for rigid and flexible ultrasonic devices" [59]. Attaching and adjusting the transducers properly by laypersons remains a challenge as domain specific knowledge (DSK) is needed for many applications.

6.4.2. Machine Learning considerations

This section discusses remaining limitations w.r.t annotations (see Section 6.4.2.1), unconsidered ML models (see Section 6.4.2.2), explainability (see Section 6.4.3.1) and the applicability of the described ML models in critical applications (see Section 6.4.3.2).

6.4.2.1. Annotations

An issue affecting all ML models is the question whether all annotations used to label the input data have been correctly and precisely obtained. In certain fields, such as classifying images of cats or dogs, labeling the data is easy and intuitive. However, this can be a much more complex issue in other fields and might also partially pose a challenge to classifications presented in this work, as precise annotation gold standards do not exist in certain cases. The following paragraphs describe potential challenges for each application.

Classifying muscle contraction and muscle fatigue states

Each subject was asked to push a button for the muscle contraction state classifications, indicating whether they currently performed a contraction or not (see Section 3.3.4). During this process, it might have happened that some subjects pushed the button either too early or too late, which would have resulted in a few wrongly annotated A-scans. However, changes in muscle contractions always occurred rather abruptly in comparison to muscle fatigue states, that changed more slowly. To this end, this work uses only sequences of A-scans from clearly distinct muscle fatigue states (i.e. the first and the last 10 seconds of each recording). Future work should include efforts to include a way of automatically annotating muscle states.

Detection of epiphyseal radius bone closure

An experienced paediatric endocrinology consultant, blinded to the age, height and weight of each patient, determined the bone age from X-ray images using the Greulich and Pyle bone age atlas for patients with a clinical indication in the epiphyseal radius bone closure study. These bone age values were also used to annotate the database. Chronological ages were used for subjects without a prior medical examination. As described in Section 4.1.2, the bone age atlas introduced by Greulich and Pyle might not be a suitable choice for all sexes and ethnicities. Hence, it might have been possible that a certain bias was introduced via wrongly annotated study subjects caused by uncertainties resulting from this way of determining bone ages. An additional potential source of bias was the gradual process of ossification, which is influenced by a variety of different factors.

Identification of hepatic steatosis and fibrosis in patients with NAFLD

There is currently no universally accepted gold standard endorsed by all members of the scientific community for liver fibrosis and liver steatosis staging or grading. Two publications from 2019 illustrate the ongoing scientific disaccord concerning the gold standard of evaluating liver fibrosis in patients with NAFLD [222, 223]. Some scientists argue that “fibrosis is a histological criteria, and biopsy is the only modality that provides histology”. They state that “all noninvasive modalities were validated using biopsy as the gold standard” and are prone to yield mislabels in the presence of significant confounders, such as obesity [222]. Other scientists have a contrary opinion and state that viewing liver biopsy as liver fibrosis assessment gold standard is an “outdated dogma”. They argue that MRE assessments should be considered superior because of “emerging data [supporting their] noninferiority to liver biopsy in terms of accuracy in fibrosis staging“. Additionally, they also mention that “although the liver biopsy complication rates appear low on an individual basis, expanding those rates to the massive and growing population in need of assessment for liver fibrosis makes it evident that the risk is inexplicable when faced with noninvasive and noninferior alternatives” [223]. A large variety of different grading and staging algorithms exists [214], which poses a further challenge to find a fitting annotation scheme. Furthermore, medical practitioners might also have introduced further bias by annotating patients not only according to their calculated grading or staging scores, but also according to other intuitive factors based on their professional experience. To address these uncertainties, this work deployed several varying annotation schemes (see Section 5.3.4).

6.4.2.2. Alternative models

Due to the large variety of existing models, this work does not include all ML models applicable for the classification of 1-D US signals. In particular, the promising technique *Transfer Learning* is not included in this work. As mentioned in Section 2.7.4.7, accessible pre-trained transfer learning models for US data were not available at the time of writing. However, as this method has become more of a focus for general TSC in recent years, future scientific research should consider this method.

6.4.2.3. Beyond binary classifications

This work focuses on the binary classification of 1-D US signals, which is performed by grouping the signals into one of two possible and distinct categories for each examined scenario. It did not examine classifications of more than two categories or the quantification of certain metrics. Previous works by other authors have shown that 1-D US signals can be used to categorize pathologies in ophthalmology [21], segment breast tissues [19, 35] or characterize soft tissues in general [30] by grouping signals in more than two categories. Hypothetically, 1-D US signals could be used to quantify certain parameters, such as exact bladder volumes of an examined patient or the exact locations of tumors. However, the methods presented in this work are not suitable to answer these types of questions as they only focus on binary TSC.

6.4.3. Regulatory considerations

An *action plan*, published by the *U.S. Food and Drug Administration* (FDA) in 2021, acknowledges that ML technologies “have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day” and highlights possible future directions to facilitate the use of those technologies in medical devices [224]. Results achieved in this work do not yet allow an immediate integration in a FDA certified medical device as further research efforts would be needed first to address regulatory issues. This section discusses remaining challenges before any official certification could be attempted.

6.4.3.1. Explainability

Recent advances in ML have ”often been achieved through increased model complexity, turning such systems into *black box* approaches and causing uncertainty regarding the way they operate and, ultimately, the way that they come to decisions”. Especially deep learning methods and ensembles are prone to be pure black boxes [225].

Solutions to this problem, such as *SHAP (SHapley Additive exPlanations)*, exist and assign each feature an importance value for a particular prediction [226]. However, it has recently been shown that “Shapley-value-based explanations for feature importance fail to serve their desired purpose in general”. For these reasons, a recent publication cautions against the usage of Shapley-value-based explanations for feature importance “except in narrowly constrained settings where they admit a clear interpretation” [227].

DRTs, such as t-SNE or PCA are helpful tools to visualize the distribution of

high-dimensional signals as a low-dimensional (most commonly 2-D or 3-D) plot. Using such techniques can drastically improve the explainability of a model by supporting hypotheses with clear visualizations. Even though those techniques are not without drawbacks, they have been extensively used in this work to substantiate initial working hypotheses with 2-D representations of high-dimensional data distributions (see Section 3.4.1, Section 3.4.3, Section 4.4.1 and Section 5.4.1). Even though these methods do not increase the explainability of certain model performances, they help tremendously to understand whether a certain hypothesis is reasonable.
Considering all factors mentioned above, the explainability of the presented models still remains to be addressed in future works in detail before the proposed solutions can be adopted in "sensitive yet critical domains, where their value could be immense, such as healthcare" [225].

6.4.3.2. Critical applications

The approaches shown in this work are not yet suitable for critical medical applications as a large database consisting of medically diverse input data is still lacking. This work examines signals of eight and 21 subjects for the muscle contraction and muscle fatigue state classifications (see Section 3.3.2), signals of 120 female subjects for epiphyseal radius bone closure detections (see Section 4.3.2) and signals of 27 patients for the classification of liver steatosis and liver fibrosis states (see Section 5.3.1). The ML models for muscle state classifications are arguably sufficient enough to be deployed in real-life scenarios. However, further research needs to be conducted for the other scenarios to allow statements regarding the universal suitability of the chosen approaches in a medical context. It would be imperative to conduct extensive clinical trials with many more subjects from a large and diverse background to take any potential bias into account before any conclusions concerning applications in critical medical applications can be drawn.

Misclassifications

It is important to correctly assess the significance of misclassifications, as their consequences vary in severity depending on the application. For example, a false negative classification of a certain steatosis grade or fibrosis stage of a patient suffering from severe liver cirrhosis is much more severe than a false negative classification for muscle contraction or muscle fatigue states in fitness applications. Hence, further research is needed to create ML models that reduce the occurrence of false negative classifications, especially in diagnosis systems as described in Chapter 4 and Chapter 5. Other scenarios, such as the one described in Chapter 3 might be able to tolerate a larger amount of false negative classifications, even though efforts should still be made to reduce those even further.

Chapter 7

Conclusions

This chapter concludes the results and observations presented and discussed in this work. Section 7.1 summarizes the findings of this work in general, while Section 7.2 portraits potential future applications. Section 7.3 elaborates on future works.

Contents

7.1. Summary

A major finding of this work is that ML models based on raw or transformed 1-D US signals can be used for a variety of biomedical binary classification tasks with sufficient degrees of accuracy and speed. Complex and very sophisticated deep learning ML models are in most cases not necessary to achieve superior results. Traditional and widely available models, such as LR or SVM are often fast and accurate alternatives for 1-D US signal classifications. These methods even outperform methods, such as 1-NN DTW, that have been deployed as

benchmark TSC algorithms for decades [81].
Pre-selection of distinct signals or feature extractions based on domain specific knowledge (DSK) are crucial parts of creating high-performing models. This work has shown that the integration of domain information, such as signal windowing according to examined soft tissue depth, removal of unnecessary signal parts or feature extraction, can reduce the time needed for training and evaluation, decrease model complexities and increase prediction performances. This work is highly relevant for future research in the field of 1-D US signal classification and parts of it have already been cited as paving "the way to accurate and fast monitoring of structural muscle features without the need for image reconstruction" [228].

7.2. Potential applications

7.2.1. Potential applications for muscle state classifications

Models used to classify muscle contraction and muscle fatigue states (see Chapter 3) can form the foundation for applications in fitness tracking or rehabilitation scenarios. These applications have in recent years gained significance in the context of the *quantified self movement*, which is a "global effort to use [...] mobile and wearable technologies to automatically obtain personal data about everyday activities" [229].
Possible future solutions include apps on mobile devices, which count the amount of muscle contractions or the extent of muscle fatigue for specific fitness sessions in a gym. One possible approach is *edge AI*, which entails the evaluation of the signals directly on a mobile device. An alternative approach would only allow for the acquisition of new testing signals on the local device, while the evaluation and / or training of the corresponding ML models would be conducted on remote *cloud computing* servers.
Additional future work might include an app on mobile devices, which measures the rehabilitation progress of patients in dedicated rehabilitation centers. For instance, similar approaches like the ones mentioned above for fitness tracking might be applied to inform about the patient's ability to perform a certain amount of muscle contractions or the patient's ability to repeat a certain exercise without suffering from incisive muscle fatigue.
Please note that those scenarios allow a certain degree of tolerance towards false positive classifications in contrast to highly critical medical diagnosis systems (see Section 6.4.3.2). Hence, the accuracies achieved in this work already suffice to serve as basis for future applications.

7.2.2. Potential applications for epiphyseal growth plate closure detection

ML models based on 1-D US signals used to detect epiphyseal growth plate closure (see Chapter 4) might find applications in scenarios that require a fast and accurate determination of bone ages. These scenarios might range from medical exams to forensic investigations ordered by authorities and carried out by executive bodies to tackle sexual exploitation of minors. Mobile devices that do not require any specific medical DSK to be operated come to mind. Providing executive bodies with such a device could allow them to tackle illegal human trafficking of underage girls and would not require any use of ionizing X-ray imaging for bone age determination.

7.2.3. Potential applications for hepatic steatosis and fibrosis detection in patients with non-alcoholic fatty liver disease

ML models used to identify hepatic steatosis and fibrosis in patients with NAFLD (see Chapter 5) could be deployed in mobile devices enabling medical practitioners, such as gastroenterologists, to assess a patient's liver speedily and accurately. Providing medical practitioners with such a device, that requires little to none specific medical DSK, could be an important cornerstone to combat the increasing prevalence of liver diseases in many countries by providing fast pre-diagnostic estimates. For instance, further medical examinations could be commissioned once such a mobile device diagnoses an early stage liver disease. As those further medical examinations often rely on expensive devices in larger healthcare centers, widely available, mobile and low-cost devices can play a vital role in decreasing medical costs overall and increasing the availability of patient care.

7.2.4. Non-destructive testing (NDT)

NDT techniques based on US are "versatile and cost-effective solutions, which have been used extensively in many industrial fields, as well as academic research tools" [18]. US imaging methods, like PWI, can be used for NDT [33] but methods based on 1-D A-scans have also been proposed [34]. The algorithms and methods proposed in this work could form the basis of future NDT applications. For instance, they could be deployed to distinguish between two or more categories of material defects or quality standards.

7.3. Future works

7.3.1. Technical aspects

1-D medical US requires the application of ultrasound gel on the examined soft tissue, which introduces a factor reducing the wearability of approaches based on this technique. Recently, a wearable single-element ultrasonic sensor "made of double-layer polyvinylidene fluoride piezoelectric polymer films with a simple and low-cost fabrication process" and including a transmitter and a receiver has been presented [58]. This work represents an important step towards future wearable setups but still depends on the application of a "medical ultrasonic gel couplant [...] between the skin surfaces and the [wearable ultrasonic sensors]" instead of integrated gel pads, which would further reduce the complexity of the required equipment [58].

The prototype of a “skin-conformal ultrasonic phased array for the monitoring of haemodynamic signals from tissues up to 14 cm beneath the skin” promises to address this challenge [59]. Further issues, such as the necessity of on-board signal processing, memory management, wireless data transmission and finding a replacement for the power supply remain to be addressed but, in general, the combination of these new hardware developments with the ML strategies developed in this work could lead to novel developments in future works and should be explored.

7.3.2. Machine Learning aspects

Transfer learning is a promising field, that has already shown good results on data similar to the data used in this work [161, 163, 164, 165, 166, 167, 230]. However, training and evaluating transfer models is a very tedious process, as this technique usually requires huge amounts of high quality input data to begin with. This would have been out of scope for this work but should be pursued in the future to show how this method compares to the presented approaches.

Additional future work is the implementation of the proposed solutions into ML models that either run directly on mobile devices (i.e. *edge AI*) or on remote cloud computing servers. For example, the library *libSVM* might be used to train and evaluate SVM models directly on *Android*, *iOS* or micro controllers [193].

A groundbreaking publication from 2016 suggests to use the 1-NN DTW algorithm, besides others, as a basic benchmark [81]. A follow-up study published in 2020, still finds that 1-NN DTW is “hard to beat and competitive with many more recently proposed alternatives” [231]. Future work looking into enhancements of this algorithm should be conducted and evaluate how those perform w.r.t. training and evaluation time and overall performances. The work at hand only includes the classical 1-NN DTW algorithm and finds it to be less competitive to other methods in terms of speed and accuracy (see Section 6.3).

7.3.3. Research fields

Even though this work provides a strong foundation for the field of 1-D US signal classifications, related research fields remain to be tackled in the future. Previous research by other authors has shown that 1-D US signals can be used to categorize pathologies in ophthalmology [21], segment breast tissues [19, 35] or characterize soft tissues in general [30]. Applying the models presented in this work to those fields should be subject of future works and could facilitate the development of smart mobile devices and smart US sensors.

Bibliography

[1] Lukas Brausch, Ruth Dirksen, Christoph Risser, Martin Schwab, Carole Stolz, Steffen Tretbar, Tilman Rohrer, and Holger Hewener. "Classification of Distal Growth Plate Ossification States of the Radius Bone Using a Dedicated Ultrasound Device and Machine Learning Techniques for Bone Age Assessments". In: *Applied Sciences* 12.7 (2022). ISSN: 2076-3417. DOI: 10.3390/app12073361.

[2] Lukas Brausch, Holger Hewener, and Paul Lukowicz. "Classifying Muscle States with One-Dimensional Radio-Frequency Signals from Single Element Ultrasound Transducers". In: *Sensors* 22.7 (2022). ISSN: 1424-8220. DOI: 10.3390/s22072789.

[3] Lukas Brausch, Holger Hewener, and Paul Lukowicz. "Towards a Wearable Low-Cost Ultrasound Device for Classification of Muscle Activity and Muscle Fatigue". In: *Proceedings of the 23rd International Symposium on Wearable Computers*. ISWC '19. London, United Kingdom: Association for Computing Machinery, (2019), pp. 20–22. ISBN: 9781450368704. DOI: 10.1145/3341163.3347749.

[4] Holger Hewener, Christoph Risser, Lukas Brausch, Tilman Rohrer, and Steffen Tretbar. "A mobile ultrasound system for majority detection". In: *2019 IEEE International Ultrasonics Symposium (IUS)*. (2019), pp. 502–505. DOI: 10.1109/ULTSYM.2019.8925868.

[5] Lukas Brausch and Holger Hewener. "Classifying muscle states with ultrasonic single element transducer data using machine learning strategies". In: *Proceedings of Meetings on Acoustics* 38.1 (2019), p. 022001. DOI: 10.1121/2.0001140.

[6] Lukas Brausch, Steffen Tretbar, and Holger Hewener. "Identification of advanced hepatic steatosis and fibrosis using ML algorithms on high-frequency ultrasound data in patients with non-alcoholic fatty liver disease". In: *2021 IEEE UFFC Latin America Ultrasonics Symposium (LAUS)*. 2021, pp. 1–4. DOI: 10.1109/LAUS53676.2021.9639128.

[7] Lukas Brausch, Holger Hewener, and Paul Lukowicz. *21 datasets of one-dimensional ultrasound raw RF data (A-scans) acquired from the calf muscles of 8 healthy volunteers.* data retrieved from OpenML.org, https://www.openml.org/d/41971. (2019).

[8] Lukas Brausch, Holger Hewener, and Paul Lukowicz. *Datasets of one-dimensional ultrasound raw RF data (A-scans) acquired from the biceps brachii muscles of 21 healthy volunteers.* data retrieved from OpenML.org, `https://www.openml.org/d/43075`. (2021).

[9] Lukas Brausch, Holger Hewener, and Paul Lukowicz. *Datasets of one-dimensional ultrasound raw RF data (A-scans) acquired from the biceps brachii muscles of a single healthy volunteer.* data retrieved from OpenML.org, `https://www.openml.org/d/43076`. (2021).

[10] Andreas Maier, Stefan Steidl, Vincent Christlein, and Joachim Hornegger. *Medical Imaging Systems: An Introductory Guide*. Vol. 11111. Springer, (2018). ISBN: 978-3-319-96519-2. DOI: `10.1007/978-3-319-96520-8`.

[11] Michael Garstang. "Long-distance, low-frequency elephant communication". In: *Journal of Comparative Physiology A* 190.10 (2004), pp. 791–805. DOI: `10.1007/s00359-004-0553-0`.

[12] Norman W. McDicken and Tom Anderson. "Basic physics of medical ultrasound". In: Jan. (2011), pp. 3–15. DOI: `10.1016/B978-0-7020-3131-1.00001-8`.

[13] Aaron Fenster and James C. Lacefield. *Ultrasound Imaging and Therapy.* Taylor & Francis, (2015). ISBN: 9781138894358. DOI: `10.1201/b18467`.

[14] Thomas L. Szabo. *Diagnostic ultrasound imaging: inside out.* Academic Press, (2004). ISBN: 978-0-12-396487-8. DOI: `10.1016/C2011-0-07261-7`.

[15] Paul Peter Urone, Roger Hinrichs, et al. "College Physics: OpenStax". In: (2018).

[16] Chang Liu, Binzhen Zhang, Chenyang Xue, Guojun Zhang, Wendong Zhang, and Yijun Cheng. "The Application of Adaptive Time Gain Compensation in an Improved Breast Ultrasound Tomography Algorithm". In: *Applied Sciences* 9.12 (2019), p. 2574. DOI: `10.3390/app9122574`.

[17] Siddhartha Sikdar, Ravi Managuli, Lixin Gong, Vijay Shamdasani, Tsuyoshi Mitake, Tetsuya Hayashi, and Yongmin Kim. "A single mediaprocessor-based programmable ultrasound system". In: *IEEE Transactions on Information Technology in Biomedicine* 7.1 (2003), pp. 64–70. DOI: `10.1109/TITB.2003.808512`.

[18] Bing Wang, Shuncong Zhong, Tung-Lik Lee, Kevin S. Fancey, and Jiawei Mi. "Non-destructive testing and evaluation of composite materials/structures: A state-of-the-art review". In: *Advances in Mechanical Engineering* 12.4 (2020), p. 1687814020913761. DOI: `10.1177/1687814020913761`.

[19] Bernard J. Ostrum, Barry B. Goldberg, and Harold J. Isard. “A-mode ultrasound differentiation of soft-tissue masses”. In: *Radiology* 88.4 (1967), pp. 745–749. DOI: `10.1148/88.4.745`.

[20] Barry B. Goldberg and J. Stauffer Lehman. “Some observations on the practical uses of A-mode ultrasound”. In: *American Journal of Roentgenology* 107.1 (1969), pp. 198–205. DOI: `10.2214/ajr.107.1.198`.

[21] William B. Trattler, Peter K. Kaiser, and Neil J. Friedman. *Review of Ophthalmology - 3rd Edition.* Elsevier Health Sciences, (2017). ISBN: 9780323390569.

[22] Christoph Amstutz, Marco Caversaccio, Jens Kowal, Richard Bächler, Lutz-Peter Nolte, Rudolf Häusler, and Martin Styner. “A-mode ultrasound–based registration in computer-aided surgery of the skull”. In: *Archives of otolaryngology–head & neck surgery* 129.12 (2003), pp. 1310–1316. DOI: `10.1001/archotol.129.12.1310`.

[23] Alpesh Patel, Viren Amin, Ronald Roberts, Doyle Wilson, and Gene Rouse. “Application of A-mode ultrasound to characterize intramuscular fat content”. In: *Review of Progress in Quantitative Nondestructive Evaluation.* Springer, (1995), pp. 1781–1788.

[24] Guilherme Ribeiro, Rafael A. de Aguiar, Rafael Penteado, Felipe D. Lisbôa, João A.G. Raimundo, Thiago Loch, Ângelo Meira, Tiago Turnes, and Fabrizio Caputo. “A-Mode Ultrasound Reliability in Fat and Muscle Thickness Measurement.” In: *Journal of Strength and Conditioning Research* (2020). DOI: `10.1519/JSC.0000000000003691`.

[25] Aiguo Han, Michal Byra, Elhamy Heba, Michael P. Andre, John W. Erdman Jr., Rohit Loomba, Claude B. Sirlin, and William D. O’Brien Jr. “Noninvasive diagnosis of nonalcoholic fatty liver disease and quantification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks”. In: *Radiology* 295.2 (2020), pp. 342–350. DOI: `10.1148/radiol.2020191160`.

[26] Naoto Shimura, Satomi Koyama, Osamu Arisaka, Mariko Imataka, Koshi Sato, and Michiko Matsuura. “Assessment of measurement of children’s bone age ultrasonically with Sunlight BonAge”. In: *Clinical Pediatric Endocrinology* 14.Supplement24 (2005), S24_17–S24_20. DOI: `10.1297/cpe.14.S24_17`.

[27] Marianna Rachmiel, Larisa Naugolni, Kineret Mazor-Aronovitch, Nira Koren-Morag, and Tzvi Bistritzer. “Bone Age Assessments by Quantitative Ultrasound (SonicBone) and Hand X-ray Based Methods are Comparable.” In: *The Israel Medical Association journal: IMAJ* 19.9 (2017), pp. 533–538.

[28] Faezeh Marzbanrad, Lisa Stroux, and Gari D. Clifford. “Cardiotocography and beyond: a review of one-dimensional Doppler ultrasound application in fetal monitoring”. In: *Physiological measurement* 39.8 (2018), 08TR01. DOI: `10.1088/1361-6579/aad4d1`.

[29] Anuja Nair, Barry D. Kuban, E. Murat Tuzcu, Paul Schoenhagen, Steven E. Nissen, and D. Geoffrey Vince. “Coronary plaque classification with intravascular ultrasound radiofrequency data analysis”. In: *Circulation* 106.17 (2002), pp. 2200–2206. DOI: `10.1161/01.cir.0000035654.18341.5e`.

[30] Ivana Despotovic, Bart Goossens, Ewout Vansteenkiste, Aleksandra Pizurica, and Wilfried Philips. “Using phase information in ultrasound RF-signals for tissue characterization”. In: *Annual work-shop on Circuits, Systems and Signal Processing (ProRISC 2008)* (2008), pp. 314–317.

[31] Pieter Kruizinga, Pim van der Meulen, Andrejs Fedjajevs, Frits Mastik, Geert Springeling, Nico de Jong, Johannes G. Bosch, and Geert Leus. “Compressive 3D ultrasound imaging using a single sensor”. In: *Science advances* 3.12 (2017), e1701423. DOI: `10.1126/sciadv.1701423`.

[32] Libertario Demi. “Practical guide to ultrasound beam forming: Beam pattern and image reconstruction analysis”. In: *Applied Sciences* 8.9 (2018), p. 1544. DOI: `10.3390/app8091544`.

[33] Léonard Le Jeune, Sébastien Robert, Eduardo Lopez Villaverde, and Claire Prada. “Plane Wave Imaging for ultrasonic non-destructive testing: Generalization to multimodal imaging”. In: *Ultrasonics* 64 (2016), pp. 128–138. DOI: `10.1016/j.ultras.2015.08.008`.

[34] Caroline Holmes, Bruce W. Drinkwater, and Paul D. Wilcox. “Post-processing of the full matrix of ultrasonic transmit–receive array data for non-destructive evaluation”. In: *NDT & e International* 38.8 (2005), pp. 701–711. DOI: `10.1016/j.ndteint.2005.04.002`.

[35] Steven Finette, Alan R. Bleier, William Swindell, and Kai Haber. “Breast Tissue Classification Using Diagnostic Ultrasound and Pattern Recognition Techniques: II. Experimental Results”. In: *Ultrasonic Imaging* 5.1 (1983). PMID: 6683020, pp. 71–86. DOI: `10.1177/016173468300500107`.

[36] G. Kossoff, W.J. Garrett, D.A. Carpenter, J. Jellins, and M.J. Dadd. “Principles and classification of soft tissues by grey scale echography”. In: *Ultrasound in Medicine & Biology* 2.2 (1976), pp. 89–105. ISSN: 0301-5629. DOI: `https://doi.org/10.1016/0301-5629(76)90017-X`.

[37] Thomas M. Deserno. “Fundamentals of medical image processing”. In: *Springer Handbook of Medical Technology*. Springer, (2011), pp. 1139–1165. DOI: `10.1007/978-3-540-74658-4_62`.

[38] Laura J. Brattain, Brian A. Telfer, Manish Dhyani, Joseph R. Grajo, and Anthony E. Samir. “Machine learning for medical ultrasound: status, methods, and future opportunities”. In: *Abdominal Radiology* 43.4 (2018), pp. 786–799. DOI: 10.1007/s00261-018-1517-0.

[39] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. “Deep learning in medical ultrasound analysis: a review”. In: *Engineering* 5.2 (2019), pp. 261–275. DOI: 10.1016/j.eng.2018.11.020.

[40] Qinghua Huang, Fan Zhang, and Xuelong Li. “Machine learning in ultrasound computer-aided diagnostic systems: a survey”. In: *BioMed research international* 2018 (2018). DOI: 10.1155/2018/5137904.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS 2012. Lake Tahoe, Nevada: Curran Associates Inc., (2012), pp. 1097–1105. DOI: 10.1145/3065386.

[42] Ruud J. G. van Sloun, Regev Cohen, and Yonina. C. Eldar. “Deep Learning in Ultrasound Imaging”. In: *Proceedings of the IEEE* 108.1 (2020), pp. 11–29. DOI: 10.1109/JPROC.2019.2932116.

[43] Dale R. Wagner, Masaru Teramoto, Trenton Judd, Joshua Gordon, Casey McPherson, and Adrianna Robison. “Comparison of A-mode and B-mode ultrasound for measurement of subcutaneous fat”. In: *Ultrasound in medicine & biology* 46.4 (2020), pp. 944–951. DOI: 10.1016/j.ultrasmedbio.2019.11.018.

[44] Oliver Amft. “How wearable computing is shaping digital health”. In: *IEEE Pervasive Computing* 17.1 (2018), pp. 92–98. DOI: 10.1109/MPRV.2018.011591067.

[45] Zheng Lou, Lili Wang, Kai Jiang, Zhongming Wei, and Guozhen Shen. “Reviews of wearable healthcare systems: Materials, devices and system integration”. In: *Materials Science and Engineering: R: Reports* 140 (2020), p. 100523. DOI: 10.1016/j.mser.2019.100523.

[46] Clarius Mobile Health. *Clarius Mobile Health*. 2022. URL: https://clarius.com/ (visited on 09/12/2022).

[47] Ba Joyce Y. Liu, Jiajun Xu, Flemming Forsberg, and Faium Ji-Bin Liu. “CMUT/CMOS-based Butterfly iQ - A Portable Personal Sonoscope”. In: *ADVANCED ULTRASOUND IN DIAGNOSIS AND THERAPY* 3 (Jan. 2019), p. 115. DOI: 10.37015/AUDT.2019.190819.

[48] Anshuman Bhuyan, Jung Woo Choe, Byung Chul Lee, Paul Cristman, Ömer Oralkan, and Butrus T. Khuri-Yakub. “Miniaturized, wearable, ultrasound probe for on-demand ultrasound screening”. In: *2011 IEEE International Ultrasonics Symposium*. IEEE. (2011), pp. 1060–1063. DOI: 10.1109/ULTSYM.2011.0260.

[49] Michail Tsakalakis and Nikolaos Bourbakis. “A wearable ultrasound multi-transducer array system for abdominal organs monitoring”. In: *13th IEEE International Conference on BioInformatics and BioEngineering*. IEEE. (2013), pp. 1–5. DOI: 10.1109/BIBE.2013.6701592.

[50] Chi-Kai Weng, Jeng-Wen Chen, Po-Yang Lee, and Chih-Chung Huang. “Implementation of a wearable ultrasound device for the overnight monitoring of tongue base deformation during obstructive sleep apnea events”. In: *Ultrasound in Medicine & Biology* 43.8 (2017), pp. 1639–1650. DOI: 10.1016/j.ultrasmedbio.2017.04.004.

[51] Sumaiya Shomaji, Parisa Dehghanzadeh, Alex Roman, Domenic Forte, Swarup Bhunia, and Soumyajit Mandal. “Early Detection of Cardiovascular Diseases Using Wearable Ultrasound Device”. In: *IEEE Consumer Electronics Magazine* 8.6 (2019), pp. 12–21. DOI: 10.1109/MCE.2019.2941350.

[52] Ang Chen, Andrew Joshua Halton, Rachel Diane Rhoades, Jayden Charles Booth, Xinhao Shi, Xiangli Bu, Ning Wu, and Junseok Chae. “Wireless Wearable Ultrasound Sensor on a Paper Substrate to Characterize Respiratory Behavior”. In: *ACS sensors* 4.4 (2019), pp. 944–952. DOI: 10.1021/acssensors.9b00043.

[53] Chonghe Wang, Xiaoshi Li, Hongjie Hu, Lin Zhang, Zhenlong Huang, Muyang Lin, Zhuorui Zhang, Zhenan Yin, Brady Huang, Hua Gong, et al. “Monitoring of the central blood pressure waveform via a conformal ultrasonic device”. In: *Nature biomedical engineering* 2.9 (2018), pp. 687–695. DOI: 10.1038/s41551-018-0287-x.

[54] Kenan Niu, Victor Sluiter, Jasper Homminga, André Sprengers, and Nico Verdonschot. “A novel ultrasound-based lower extremity motion tracking system”. In: *Intelligent Orthopaedics*. Springer, (2018), pp. 131–142. DOI: 10.1007/978-981-13-1396-7_11.

[55] IntelaMetrix. *BodyMetrix BX2000 receives FDA 510(k) clearance.* 2009. URL: https://www.accessdata.fda.gov/cdrh_docs/pdf8/K082147.pdf (visited on 04/01/2021).

[56] Renata M. Bielemann, Maria Cristina Gonzalez, Thiago Gonzalez Barbosa-Silva, Silvana Paiva Orlandi, Mariana Otero Xavier, Rafaela Bülow Bergmann, Maria Cecília Formoso Assunção, et al. “Estimation of body fat in adults using a portable A-mode ultrasound”. In: *Nutrition* 32.4 (2016), pp. 441–446. DOI: 10.1016/j.nut.2015.10.009.

[57] Monica Miclos-Balica, Paul Muntean, Falk Schick, Horia G. Haragus, Bogdan Glisici, Vasile Pupazan, Adrian Neagu, and Monica Neagu. “Reliability of body composition assessment using A-mode ultrasound in a heterogeneous sample”. In: *European Journal of Clinical Nutrition* (2020), pp. 1–8. DOI: 10.1038/s41430-020-00743-y.

[58] Ibrahim AlMohimeed and Yuu Ono. “Ultrasound measurement of skeletal muscle contractile parameters using flexible and wearable single-element ultrasonic sensor”. In: *Sensors* 20.13 (2020), p. 3616. DOI: 10.3390/s20133616.

[59] Chonghe Wang, Baiyan Qi, Muyang Lin, Zhuorui Zhang, Mitsutoshi Makihata, Boyu Liu, Sai Zhou, Yi-hsi Huang, Hongjie Hu, Yue Gu, et al. “Continuous monitoring of deep-tissue haemodynamics with stretchable ultrasonic phased arrays”. In: *Nature Biomedical Engineering* 5.7 (2021), pp. 749–758.

[60] Reza Arghandeh and Yuxun Zhou. *Big data application in power systems*. Elsevier, (2017). ISBN: 978-0-12-811968-6. DOI: 10.1016/C2016-0-00194-8.

[61] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. “Deep learning for time series classification: a review”. In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963. DOI: 10.1007/s10618-019-00619-1.

[62] Marc Wenninger, Sebastian P Bayerl, Jochen Schmidt, and Korbinian Riedhammer. “Timage–A Robust Time Series Classification Pipeline”. In: *International Conference on Artificial Neural Networks*. Springer. (2019), pp. 450–461.

[63] Theophano Mitsa. *Temporal data mining*. CRC Press, (2010). ISBN: 978-1-4200-8976-9.

[64] Yi-Wen Liu. “Hilbert transform and applications”. In: *Fourier Transform Applications* (2012), pp. 291–300.

[65] Alain de Cheveigné and Israel Nelken. “Filters: when, why, and how (not) to use them”. In: *Neuron* 102.2 (2019), pp. 280–293. DOI: 10.1016/j.neuron.2019.02.039.

[66] Marília Barandas, Duarte Folgado, Letícia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. “Tsfel: Time series feature extraction library”. In: *SoftwareX* 11 (2020), p. 100456. DOI: 10.1016/j.softx.2020.100456.

[67] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*. Vol. 208. Springer, (2001).

[68] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, et al. “Array programming with NumPy”. In: *Nature* 585 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[69] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[70] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. “Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)”. In: *Neurocomputing* 307 (2018), pp. 72–77. DOI: 10.1016/j.neucom.2018.03.067.

[71] William L. Briggs and Van Emden Henson. *The DFT: an owner’s manual for the discrete Fourier transform.* SIAM, (1995). DOI: 10.1137/1.9781611971514.

[72] Sabur Ajibola Alim and Nahrul Khair Alang Rashid. “Some commonly used speech feature extraction algorithms”. In: IntechOpen, Dec. (2018). ISBN: 978-1-78984-702-4. DOI: 10.5772/intechopen.80419.

[73] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. “The timbre toolbox: Extracting audio descriptors from musical signals”. In: *The Journal of the Acoustical Society of America* 130.5 (2011), pp. 2902–2916. DOI: 10.1121/1.3642604.

[74] Shijo M. Joseph and Anto P. Babu. “Wavelet energy based voice activity detection and adaptive thresholding for efficient speech coding”. In: *International Journal of Speech Technology* 19.3 (2016), pp. 537–550. DOI: 10.1007/s10772-014-9240-x.

[75] Banfu Yan, Ayaho Miyamoto, and Eugen Brühwiler. “Wavelet transform-based modal parameter identification considering uncertainty”. In: *Journal of Sound and Vibration* 291.1-2 (2006), pp. 285–301. DOI: 10.1016/J.JSV.2005.06.005.

[76] Abdeslam Serroukh, Andrew T. Walden, and Donald B. Percival. “Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series”. In: *Journal of the American Statistical Association* 95.449 (2000), pp. 184–196. DOI: 10.2307/2669537.

[77] Christopher M. Bishop. *Pattern recognition and machine learning.* Springer, (2006). ISBN: 0387310738. DOI: 10.1117/1.2819119.

[78] Qiang Yang and Xindong Wu. “10 challenging problems in data mining research”. In: *International Journal of Information Technology & Decision Making* 5.04 (2006), pp. 597–604. DOI: 10.1142/S0219622006002258.

[79] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. “The UCR time series archive”. In: *IEEE/CAA Journal of Automatica Sinica* 6.6 (2019), pp. 1293–1305.

[80] Amaia Abanda, Usue Mori, and Jose A. Lozano. “A review on distance based time series classification”. In: *Data Mining and Knowledge Discovery* 33.2 (2019), pp. 378–412. DOI: 10.1007/s10618-018-0596-4.

[81] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. “The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances”. In: *Data mining and knowledge discovery* 31.3 (2017), pp. 606–660. DOI: 10.1007/s10618-016-0483-9.

[82] Patrick Schäfer. “The BOSS is concerned with time series classification in the presence of noise”. In: *Data Mining and Knowledge Discovery* 29.6 (2015), pp. 1505–1530. DOI: 10.1007/s10618-014-0377-7.

[83] Angus Dempster, François Petitjean, and Geoffrey I. Webb. “ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels”. In: *Data Mining and Knowledge Discovery* 34.5 (2020), pp. 1454–1495. DOI: 10.1007/s10618-020-00701-z.

[84] Patrick Schäfer and Ulf Leser. “Fast and accurate time series classification with weasel”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* (2017), pp. 637–646. DOI: 10.1145/3132847.3132980.

[85] Lexiang Ye and Eamonn Keogh. “Time series shapelets: a new primitive for data mining”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.* (2009), pp. 947–956. DOI: 10.1145/1557019.1557122.

[86] Jason Lines, Luke M. Davis, Jon Hills, and Anthony Bagnall. “A shapelet transform for time series classification”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* (2012), pp. 289–297. DOI: 10.1145/2339530.2339579.

[87] Aaron Bostrom and Anthony Bagnall. “Binary shapelet transform for multiclass time series classification”. In: *International conference on big data analytics and knowledge discovery.* Springer. (2015), pp. 257–269. DOI: 10.1007/978-3-319-22729-0_20.

[88] Houtao Deng, George Runger, Eugene Tuv, and Martyanov Vladimir. “A time series forest for classification and feature extraction”. In: *Information Sciences* 239 (2013), pp. 142–153. DOI: 10.1016/j.ins.2013.02.030.

[89] Matthew Middlehurst, James Large, and Anthony Bagnall. “The canonical interval forest (CIF) classifier for time series classification”. In: *arXiv preprint arXiv:2008.09172* (2020).

[90] Mustafa Gokce Baydogan and George Runger. “Time series representation and similarity based on local autopatterns”. In: *Data Mining and Knowledge Discovery* 30.2 (2016), pp. 476–509. DOI: 10.1007/s10618-015-0425-y.

[91] Michael Flynn, James Large, and Tony Bagnall. "The contract random interval spectral ensemble (c-RISE): the effect of contracting a classifier on accuracy". In: *International Conference on Hybrid Artificial Intelligence Systems.* Springer. (2019), pp. 381–392. DOI: 10.1007/978-3-030-29859-3_33.

[92] Jason Lines and Anthony Bagnall. "Time series classification with ensembles of elastic distance measures". In: *Data Mining and Knowledge Discovery* 29.3 (2015), pp. 565–592. DOI: 10.1007/s10618-014-0361-2.

[93] Benjamin Lucas, Ahmed Shifaz, Charlotte Pelletier, Lachlan O'Neill, Nayyar Zaidi, Bart Goethals, François Petitjean, and Geoffrey I. Webb. "Proximity forest: an effective and scalable distance-based classifier for time series". In: *Data Mining and Knowledge Discovery* 33.3 (2019), pp. 607–635. DOI: 10.1007/s10618-019-00617-3.

[94] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. "Time-series classification with COTE: the collective of transformation-based ensembles". In: *IEEE Transactions on Knowledge and Data Engineering* 27.9 (2015), pp. 2522–2535. DOI: 10.1109/TKDE.2015.2416723.

[95] Jason Lines, Sarah Taylor, and Anthony Bagnall. "Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification". In: *2016 IEEE 16th international conference on data mining (ICDM).* IEEE. (2016), pp. 1041–1046. DOI: 10.1109/ICDM.2016.0133.

[96] Jason Lines, Sarah Taylor, and Anthony Bagnall. "Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles". In: *ACM Transactions on Knowledge Discovery from Data* 12.5 (2018). DOI: 10.1145/3182382.

[97] Ahmed Shifaz, Charlotte Pelletier, François Petitjean, and Geoffrey I. Webb. "TS-CHIEF: a scalable and accurate forest algorithm for time series classification". In: *Data Mining and Knowledge Discovery* 34.3 (2020), pp. 742–775. DOI: 10.1007/s10618-020-00679-8.

[98] Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. "catch22: Canonical time-series characteristics". In: *Data Mining and Knowledge Discovery* 33.6 (2019), pp. 1821–1852. DOI: 10.1007/s10618-019-00647-x.

[99] Matthew Middlehurst, James Large, Gavin Cawley, and Anthony Bagnall. "The Temporal Dictionary Ensemble (TDE) Classifier for Time Series Classification". In: *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.* (2020).

[100] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. "Inceptiontime: Finding AlexNet for time series classification". In: *Data Mining and Knowledge Discovery* 34.6 (2020), pp. 1936–1962. DOI: `10.1007/s10618-020-00710-y`.

[101] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* (2016), pp. 785–794. DOI: `10.1145/2939672.2939785`.

[102] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[103] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. "CatBoost: unbiased boosting with categorical features". In: *arXiv preprint arXiv:1706.09516* (2017).

[104] Martin Kropf, Dieter Hayn, Daniel Morris, Aravind-Kumar Radhakrishnan, Evgeny Belyavskiy, Athanasios Frydas, Elisabeth Pieske-Kraigher, Burkert Pieske, and Günter Schreier. "Cardiac anomaly detection based on time and frequency domain features using tree-based classifiers". In: *Physiological measurement* 39.11 (2018), p. 114001. DOI: `10.1088/1361-6579/aae13e`.

[105] Hong Zeng, Chen Yang, Hua Zhang, Zhenhua Wu, Jiaming Zhang, Guojun Dai, Fabio Babiloni, and Wanzeng Kong. "A lightGBM-based EEG analysis method for driver mental states classification". In: *Computational intelligence and neuroscience* 2019 (2019). DOI: `10.1155/2019/3761203`.

[106] John T. Hancock and Taghi M. Khoshgoftaar. "CatBoost for big data: an interdisciplinary review". In: *Journal of big Data* 7.1 (2020), pp. 1–45. DOI: `10.1186/s40537-020-00369-8`.

[107] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117. DOI: `10.1016/j.neunet.2014.09.003`.

[108] Friedhelm Schwenker, Christian Dietrich, Hans A. Kestler, Klaus Riede, and Günther Palm. "Radial basis function neural networks and temporal fusion for the classification of bioacoustic time series". In: *Neurocomputing* 51 (2003), pp. 265–275. DOI: `10.1016/S0925-2312(02)00621-5`.

[109] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling." In: *CoRR* abs/1803.01271 (2018).

[110] Zhiguang Wang, Weizhong Yan, and Tim Oates. "Time series classification from scratch with deep neural networks: A strong baseline". In: *2017 International joint conference on neural networks (IJCNN)*. IEEE. (2017), pp. 1578–1585.

[111] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. "Rethinking 1d-cnn for time series classification: A stronger baseline". In: *arXiv preprint arXiv:2002.10061* (2020).

[112] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. "MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification". In: *arXiv preprint arXiv:2012.08791* (2020). DOI: 10.1145/3447548.3467231.

[113] Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I. Webb. "MultiRocket: Effective summary statistics for convolutional outputs in time series classification". In: *arXiv preprint arXiv:2102.00457* (2021).

[114] Jean-Pierre Eckmann, Sylvie O. Kamphorst, and D. Ruelle. "Recurrence Plots of Dynamical Systems". In: *Europhysics Letters (EPL)* 4.9 (1987), pp. 973–977. DOI: 10.1209/0295-5075/4/9/004.

[115] Nima Hatami, Yann Gavet, and Johan Debayle. "Classification of time-series images using deep convolutional neural networks". In: *Tenth international conference on machine vision (ICMV 2017)*. Vol. 10696. International Society for Optics and Photonics. (2018), 106960Y. DOI: 10.1117/12.2309486.

[116] Enrique Garcia-Ceja, Md Zia Uddin, and Jim Torresen. "Classification of recurrence plots' distance matrices with a convolutional neural network for activity recognition". In: *Procedia computer science* 130 (2018), pp. 157–163. DOI: 10.1016/j.procs.2018.04.025.

[117] Ye Zhang, Yi Hou, Shilin Zhou, and Kewei Ouyang. "Encoding time series as multi-scale signed recurrence plots for classification using fully convolutional networks". In: *Sensors* 20.14 (2020), p. 3818. DOI: 10.3390/s20143818.

[118] Zhiguang Wang and Tim Oates. "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks". In: *Workshops at the twenty-ninth AAAI conference on artificial intelligence*. Vol. 1. (2015).

[119] Mark French and Rod Handy. "Spectrograms: turning signals into pictures". In: *Journal of engineering technology* 24.1 (2007), p. 32.

[120] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[121] Ben D. Fulcher and Nick S. Jones. "hctsa: A computational framework for automated time-series phenotyping using massive feature extraction". In: *Cell systems* 5.5 (2017), pp. 527–531. DOI: `10.1016/j.cels.2017.10.001`.

[122] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[123] Dalwinder Singh and Birmohan Singh. "Investigating the impact of data normalization on classification performance". In: *Applied Soft Computing* 97 (2020), p. 105524. DOI: `10.1016/j.asoc.2019.105524`.

[124] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[125] Ganesh Naik. *Biomedical Signal Processing.* Springer, (2020). ISBN: 978-981-13-9099-9. DOI: `10.1007/978-981-13-9097-5`.

[126] Aliaksei Sandryhaila and Jose MF Moura. "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure". In: *IEEE Signal Processing Magazine* 31.5 (2014), pp. 80–90. DOI: `10.1109/MSP.2014.2329213`.

[127] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. "Graph signal processing: Overview, challenges, and applications". In: *Proceedings of the IEEE* 106.5 (2018), pp. 808–828. DOI: `10.1109/JPROC.2018.2820126`.

[128] Russell Stuart and Norvig Peter. *Artificial intelligence-a modern approach 3rd edition.* (2016).

[129] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. "Rethinking bias-variance trade-off for generalization of neural networks". In: *International Conference on Machine Learning.* PMLR. (2020), pp. 10767–10777.

[130] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. (2015).

[131] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc., (2019), pp. 8024–8035.

[132] François Chollet et al. *Keras.* `https://keras.io`. (2015).

[133] John D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: `10.1109/MCSE.2007.55`.

[134] Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J Király. "sktime: A unified interface for machine learning with time series". In: *arXiv preprint arXiv:1909.07872* (2019).

[135] Wannes Meert, Kilian Hendrickx, and Toon Van Craenendonck. *wannesm/dtaidistance v2.0.0*. Version v2.0.0. Aug. 2020. DOI: `10.5281/zenodo.3981067`. URL: `https://doi.org/10.5281/zenodo.3981067`.

[136] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[137] John P. Cunningham and Zoubin Ghahramani. "Linear dimensionality reduction: Survey, insights, and generalizations". In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 2859–2900.

[138] Richard Bellman and Robert Kalaba. "On adaptive control processes". In: *IRE Transactions on Automatic Control* 4.2 (1959), pp. 1–9. DOI: `10.1109/TAC.1959.1104847`.

[139] Pavel Senin. "Dynamic time warping algorithm review". In: *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855.1-23 (2008), p. 40.

[140] Wei-Yin Loh. "Fifty years of classification and regression trees". In: *International Statistical Review* 82.3 (2014), pp. 329–348. DOI: `10.1111/insr.12016`.

[141] Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, Safaa Abdel Raouf, Samar Darweesh, Mohamad Soliman, Mohamed Elhefnawi, Mohamed El-Adawy, and Mahmoud ElHefnawi. "Accurate prediction of advanced liver fibrosis using the decision tree learning algorithm in chronic hepatitis C Egyptian patients". In: *Gastroenterology research and practice* 2016 (2016). DOI: `10.1155/2016/2636390`.

[142] Leo Breiman. *Arcing the edge*. Tech. rep. Statistics Department, University of California, 1997.

[143] Bulat Ibragimov and Gleb Gusev. "Minimal variance sampling in stochastic gradient boosting". In: *arXiv preprint arXiv:1910.13204* (2019).

[144] Alexey Ya Chervonenkis. "Early history of support vector machines". In: *Empirical Inference*. Springer, (2013), pp. 13–20. DOI: `10.1007/978-3-642-41136-6_3`.

[145] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. "A comprehensive survey on support vector machine classification: Applications, challenges and trends". In: *Neurocomputing* 408 (2020), pp. 189–215. DOI: `10.1016/j.neucom.2019.10.118`.

[146] Tomasz Szandała. “Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks”. In: *Bio-inspired Neurocomputing.* Springer, (2021), pp. 203–224. DOI: 10.1007/978-981-15-5495-7_11.

[147] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems.* (2017), pp. 5998–6008.

[148] Minghao Liu, Shengqi Ren, Siyuan Ma, Jiahui Jiao, Yizhou Chen, Zhiguang Wang, and Wei Song. “Gated Transformer Networks for Multivariate Time Series Classification”. In: *arXiv preprint arXiv:2103.14438* (2021).

[149] Tarek Allam Jr. and Jason D. McEwen. “Paying Attention to Astronomical Transients: Photometric Classification with the Time-Series Transformer”. In: *arXiv preprint arXiv:2105.06178* (2021).

[150] Juan José Rodriguez, Ludmila I. Kuncheva, and Carlos J. Alonso. “Rotation forest: A new classifier ensemble method”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.10 (2006), pp. 1619–1630. DOI: 10.1109/TPAMI.2006.211.

[151] Peter Sykacek and Stephen J. Roberts. “Bayesian time series classification”. In: *Advances in Neural Information Processing Systems* 14 (2002), p. 937.

[152] Kristian Gundersen, Guttorm Alendal, Anna Oleynik, and Nello Blaser. “Binary time series classification with bayesian convolutional neural networks when monitoring for marine gas discharges”. In: *Algorithms* 13.6 (2020), p. 145.

[153] Patrick Schäfer. “Scalable time series classification”. In: *Data Mining and Knowledge Discovery* 30.5 (2016), pp. 1273–1298. DOI: 10.1007/s10618-015-0441-y.

[154] Matthew Middlehurst, William Vickers, and Anthony Bagnall. “Scalable dictionary classifiers for time series classification”. In: *International Conference on Intelligent Data Engineering and Automated Learning.* Springer. (2019), pp. 11–19.

[155] James Large, Anthony Bagnall, Simon Malinowski, and Romain Tavenard. “On time series classification with dictionary-based classifiers”. In: *Intelligent Data Analysis* 23.5 (2019), pp. 1073–1089. DOI: 10.3233/IDA-184333.

[156] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. “Deep neural networks as gaussian processes”. In: *arXiv preprint arXiv:1711.00165* (2017).

[157] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. "When Gaussian process meets big data: A review of scalable GPs". In: *IEEE transactions on neural networks and learning systems* 31.11 (2020), pp. 4405–4423.

[158] Zachary C. Lipton, John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning". In: *arXiv preprint arXiv:1506.00019* (2015).

[159] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. "Deep transformer models for time series forecasting: The influenza prevalence case". In: *arXiv preprint arXiv:2001.08317* (2020).

[160] Anthony Bagnall, Aaron Bostrom, Gavin C. Cawley, M. Flynn, J. Large, and Jason Lines. "Is rotation forest the best classifier for problems with continuous features?" In: *arXiv preprint arXiv:1809.06705* (2018).

[161] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. "Transfer learning for time series classification". In: *2018 IEEE international conference on big data (Big Data)*. IEEE. (2018), pp. 1367–1376.

[162] Pankaj Malhotra, Vishnu Tv, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. "TimeNet: Pre-trained deep recurrent neural network for time series classification". In: *arXiv preprint arXiv:1706.08838* (2017).

[163] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Xinyu Huang, and Marcin Grzegorzek. "Deep transfer learning for time series data based on sensor modality classification". In: *Sensors* 20.15 (2020), p. 4271. DOI: 10.3390/s20154271.

[164] Neeti Wagle and Eric W. Frew. "Transfer learning for dynamic RF environments". In: *2012 American Control Conference (ACC)*. IEEE. (2012), pp. 1406–1411. DOI: 10.1109/ACC.2012.6315333.

[165] Shekoofeh Azizi, Parvin Mousavi, Pingkun Yan, Amir Tahmasebi, Jin Tae Kwak, Sheng Xu, Baris Turkbey, Peter Choyke, Peter Pinto, Bradford Wood, et al. "Transfer learning from RF to B-mode temporal enhanced ultrasound features for prostate cancer detection". In: *International journal of computer assisted radiology and surgery* 12.7 (2017), pp. 1111–1121. DOI: 10.1007/s11548-017-1573-x.

[166] Xueli Wang, Yufeng Zhang, Hongxin Zhang, Yixuan Li, and Xiaofeng Wei. "Radio frequency signal identification using transfer learning based on LSTM". In: *Circuits, Systems, and Signal Processing* 39.11 (2020), pp. 5514–5528. DOI: 10.1007/s00034-020-01417-7.

[167] Scott Kuzdeba, Josh Robinson, and Joseph Carmack. "Transfer learning with radio frequency signals". In: *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE. (2021), pp. 1–9. DOI: 10.1109/CCNC49032.2021.9369550.

[168] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), pp. 1–13. DOI: 10.1186/s12864-019-6413-7.

[169] Nobutaka Mitsuhashi, Kaori Fujieda, Takuro Tamura, Shoko Kawamoto, Toshihisa Takagi, and Kousaku Okubo. "BodyParts3D: 3D structure database for anatomical concepts". In: *Nucleic acids research* 37.suppl_1 (2009), pp. D782–D785.

[170] Walter Herzog, Krysta Powers, Kaleena Johnston, and Mike Duvall. "A new paradigm for muscle contraction". In: *Frontiers in physiology* 6 (2015), p. 174. DOI: 10.3389/fphys.2015.00174.

[171] Simon C. Gandevia. "Spinal and supraspinal factors in human muscle fatigue". In: *Physiological reviews* (2001). DOI: 10.1152/physrev.2001.81.4.1725.

[172] Jing-jing Wan, Zhen Qin, Peng-yuan Wang, Yang Sun, and Xia Liu. "Muscle fatigue: general understanding and treatment". In: *Experimental & molecular medicine* 49.10 (2017), e384–e384. DOI: 10.1038/emm.2017.194.

[173] Néstor J. Jarque-Bou, Joaquín L. Sancho-Bru, and Margarita Vergara. "A Systematic Review of EMG Applications for the Characterization of Forearm and Hand Muscle Activity during Activities of Daily Living: Results, Challenges, and Open Issues". In: *Sensors* 21.9 (2021), p. 3035. DOI: 10.3390/s21093035.

[174] Hari Prasanth, Miroslav Caban, Urs Keller, Grégoire Courtine, Auke Ijspeert, Heike Vallery, and Joachim Von Zitzewitz. "Wearable sensor-based real-time gait detection: a systematic review". In: *Sensors* 21.8 (2021), p. 2727. DOI: 10.3390/s21082727.

[175] Oliver Amft, Holger Junker, Paul Lukowicz, Gerhard Troster, and Corina Schuster. "Sensing muscle activities with body-worn sensors". In: *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06).* IEEE. (2006), 4–pp. DOI: 10.1109/BSN.2006.48.

[176] Alok Prakash, Neeraj Sharma, and Shiru Sharma. "Novel force myography sensor to measure muscle contractions for controlling hand prostheses". In: *Instrumentation Science & Technology* 48.1 (2020), pp. 43–62. DOI: 10.1080/10739149.2019.1655441.

[177] Jan Meyer, Paul Lukowicz, and Gerhard Troster. "Textile pressure sensor for muscle activity and motion detection". In: *2006 10th IEEE International Symposium on Wearable Computers.* IEEE. (2006), pp. 69–72. DOI: 10.1109/ISWC.2006.286346.

[178] Bo Zhou, Mathias Sundholm, Jingyuan Cheng, Heber Cruz, and Paul Lukowicz. "Measuring muscle activities during gym exercises with textile pressure mapping sensors". In: *Pervasive and Mobile Computing* 38 (2017), pp. 331–345. DOI: 10.1016/j.pmcj.2016.08.015.

[179] Yadong Xu, Ganggang Zhao, Liang Zhu, Qihui Fei, Zhe Zhang, Zanyu Chen, Fufei An, Yangyang Chen, Yun Ling, Peijun Guo, et al. "Pencil–paper on-skin electronics". In: *Proceedings of the National Academy of Sciences* 117.31 (2020), pp. 18292–18301. DOI: 10.1073/pnas.2008422117.

[180] Bas Van Hooren, Panayiotis Teratsias, and Emma F. Hodson-Tole. "Ultrasound imaging to assess skeletal muscle architecture during movements: a systematic review of methods, reliability, and challenges". In: *Journal of Applied Physiology* 128.4 (2020), pp. 978–999. DOI: 10.1152/japplphysiol.00835.2019.

[181] Jing-Yi Guo, Yong-Ping Zheng, Qing-Hua Huang, Xin Chen, et al. "Dynamic monitoring of forearm muscles using one-dimensional sonomyography system". In: (2008). DOI: 10.1682/jrrd.2007.02.0026.

[182] Nalinda Hettiarachchi, Zhaojie Ju, and Honghai Liu. "A new wearable ultrasound muscle activity sensing system for dexterous prosthetic control". In: *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE. (2015), pp. 1415–1420. DOI: 10.1109/SMC.2015.251.

[183] Yuefeng Li, Keshi He, Xueli Sun, and Honghai Liu. "Human-machine interface based on multi-channel single-element ultrasound transducers: A preliminary study". In: *2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom)*. IEEE. (2016), pp. 1–6. DOI: 10.1109/HealthCom.2016.7749483.

[184] Xueli Sun, Xingchen Yang, Xiangyang Zhu, and Honghai Liu. "Dual-frequency ultrasound transducers for the detection of morphological changes of deep-layered muscles". In: *IEEE Sensors Journal* 18.4 (2017), pp. 1373–1383. DOI: 10.1109/JSEN.2017.2778243.

[185] Xingchen Yang, Xueli Sun, Dalin Zhou, Yuefeng Li, and Honghai Liu. "Towards wearable A-mode ultrasound sensing for real-time finger motion recognition". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.6 (2018), pp. 1199–1208. DOI: 10.1109/TNSRE.2018.2829913.

[186] Yu Zhou, Jingbiao Liu, Jia Zeng, Kairu Li, and Honghai Liu. "Bio-signal based elbow angle and torque simultaneous prediction during isokinetic contraction". In: *Science China Technological Sciences* 62.1 (2019), pp. 21–30. DOI: 10.1007/s11431-018-9354-5.

[187] Xingchen Yang, Jipeng Yan, and Honghai Liu. "Comparative analysis of wearable a-mode ultrasound and SEMG for muscle-computer interface". In: *IEEE Transactions on Biomedical Engineering* 67.9 (2019), pp. 2434–2442. DOI: 10.1109/TBME.2019.2962499.

[188] Xingchen Yang, Zhenfeng Chen, Nalinda Hettiarachchi, Jipeng Yan, and Honghai Liu. “A wearable ultrasound system for sensing muscular morphological deformations”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2019). DOI: 10.1109/TSMC.2019.2924984.

[189] Wei Xia, Yu Zhou, Xingchen Yang, Keshi He, and Honghai Liu. “Toward portable hybrid surface electromyography/a-mode ultrasound sensing for human–machine interface”. In: *IEEE Sensors Journal* 19.13 (2019), pp. 5219–5228. DOI: 10.1109/JSEN.2019.2903532.

[190] Xingchen Yang, Jipeng Yan, Zhenfeng Chen, Han Ding, and Honghai Liu. “A proportional pattern recognition control scheme for wearable a-mode ultrasound sensing”. In: *IEEE Transactions on Industrial Electronics* 67.1 (2019), pp. 800–808. DOI: 10.1109/TIE.2019.2898614.

[191] Jipeng Yan, Xingchen Yang, Zhenfeng Chen, and Honghai Liu. “Dynamically Characterizing Skeletal Muscles via Acoustic Non-linearity Parameter: In Vivo Assessment for Upper Arms”. In: *Ultrasound in medicine & biology* 46.2 (2020), pp. 315–324. DOI: 10.1016/j.ultrasmedbio.2019.08.007.

[192] Xueli Sun, Yuefeng Li, and Honghai Liu. “Muscle fatigue assessment using one-channel single-element ultrasound transducer”. In: *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. (2017), pp. 122–125. DOI: 10.1109/NER.2017.8008307.

[193] Chih-Chung Chang and Chih-Jen Lin. “LIBSVM: a library for support vector machines”. In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27. DOI: 10.1145/1961189.1961199.

[194] Fabio Sarto, Jörg Spörri, Daniel P. Fitze, Jonathan I. Quinlan, Marco V. Narici, and Martino V. Franchi. “Implementing ultrasound imaging for the assessment of muscle and tendon properties in elite sports: Practical aspects, methodological considerations and future directions”. In: *Sports Medicine* (2021), pp. 1–20. DOI: 10.1007/s40279-021-01436-7.

[195] Federica Cavallo, Angelika Mohn, Francesco Chiarelli, and Cosimo Giannini. “Evaluation of Bone Age in Children: A Mini-Review”. In: *Frontiers in Pediatrics* 9 (2021), p. 21. ISSN: 2296-2360. DOI: 10.3389/fped.2021.580314.

[196] Pieter J. J. Sauer, Alf Nicholson, David Neubauer, et al. *Age determination in asylum seekers: physicians should not be implicated.* (2016). DOI: 10.1007/s00431-015-2628-z.

[197] Andrew T. Pennock, James D. Bomar, and John D. Manning. “The creation and validation of a knee bone age atlas utilizing MRI”. In: *JBJS* 100.4 (2018), e20. DOI: 10.2106/JBJS.17.00693.

[198] Khalaf Alshamrani, Fabrizio Messina, and Amaka C. Offiah. “Is the Greulich and Pyle atlas applicable to all ethnicities? A systematic review and meta-analysis”. In: *European radiology* 29.6 (2019), pp. 2910–2923. DOI: `10.1007/s00330-018-5792-5`.

[199] Ana L. Creo and W. Frederick Schwenk. “Bone age: a handy tool for pediatric providers”. In: *Pediatrics* 140.6 (2017). DOI: `10.1542/peds.2017-1486`.

[200] ADMV Castriota-Scanderbeg and V. De Micheli. “Ultrasound of femoral head cartilage: a new method of assessing bone age”. In: *Skeletal radiology* 24.3 (1995), pp. 197–200. DOI: `10.1007/BF00228922`.

[201] Alessandro Castriota-Scanderbeg, Michele C. Sacco, Leonardo Emberti-Gialloreti, and Lucio Fraracci. “Skeletal age assessment in children and young adults: comparison between a newly developed sonographic method and conventional methods”. In: *Skeletal radiology* 27.5 (1998), pp. 271–277. DOI: `10.1007/s002560050380`.

[202] Hans-Joachim Mentzel, Claudia Vilser, Marcus Eulenstein, Tseela Schwartz, Susanna Vogt, Joachim Böttcher, Irit Yaniv, Liat Tsoref, Eberhard Kauf, and Werner A. Kaiser. “Assessment of skeletal age at the wrist in children with a new ultrasound device”. In: *Pediatric radiology* 35.4 (2005), pp. 429–433. DOI: `10.1007/s00247-004-1385-3`.

[203] Hans-Joachim Mentzel, Susanna Vogt, Claudia Vilser, Tseela Schwartz, Marcus Eulenstein, Joachim Böttcher, Liat Tsoref, Eberhard Kauf, and Werner A. Kaiser. “Abschätzung des Knochenalters mit einer neuen Ultraschallmethode”. In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren.* Vol. 177. 12. © Georg Thieme Verlag KG Stuttgart · New York. (2005), pp. 1699–1705. DOI: `10.1055/s-2005-858764`.

[204] Pascal Laugier, Maryline Talmant, and Pham Thien-Ly. “Quo vadis, ultrasonics of bone? Present state and future trends”. In: *Archives of Acoustics* 33.4 (2008), pp. 553–564. DOI: `10.1121/1.2934866`.

[205] Khalid M. Khan, Bradley S. Miller, Eric Hoggard, Arif Somani, and Kyriakie Sarafoglou. “Application of ultrasound for bone age estimation in clinical practice”. In: *The Journal of pediatrics* 154.2 (2009), pp. 243–247. DOI: `10.1016/j.jpeds.2008.08.018`.

[206] Marianna Rachmiel, Larisa Naugolani, Kineret Mazor-Aronovitch, A Levin, Nira Koren-Morag, and Tzvi Bistritzer. “Bone age assessment by a novel quantitative ultrasound based device (SonicBone), is comparable to the conventional Greulich and Pyle method”. In: *Horm Res Pediatr* 80.Suppl 1 (2013), p. 35.

[207] Oguzhan Ekizoglu, Ali Er, Asli Dilara Buyuktoka, Mustafa Bozdag, Gokce Karaman, Negahnaz Moghaddam, and Silke Grabherr. "Ultrasonographic assessment of ossification of the distal radial epiphysis for estimating forensic age". In: *International Journal of Legal Medicine* 135.4 (2021), pp. 1573–1580.

[208] Zhiwei Chen, Wenqiang Luo, Qi Zhang, Baiying Lei, Tianfu Wang, Zhong Chen, Yuan Fu, Peidong Guo, Changchuan Li, Teng Ma, et al. "Osteoporosis Diagnosis Based on Ultrasound Radio Frequency Signal via Multi-channel Convolutional Neural Network". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 832–835. DOI: 10.1109/EMBC46164.2021.9629546.

[209] Takuya Ishimoto, Ryoichi Suetoshi, Dorian Cretin, Koji Hagihara, Jun Hashimoto, Akio Kobayashi, and Takayoshi Nakano. "Quantitative ultrasound (QUS) axial transmission method reflects anisotropy in micro-arrangement of apatite crystallites in human long bones: A study with 3-MHz-frequency ultrasound". In: *Bone* 127 (2019), pp. 82–90. DOI: 10.1016/j.bone.2019.05.034.

[210] Fátima Baptista, Lurdes M Rebocho, Graça Cardadeiro, Vera Zymbal, and Nicoletta Rosati. "Sex-and maturity-related differences in cortical bone at the distal radius and midshaft tibia evaluated by quantitative ultrasonography". In: *Ultrasound in medicine & biology* 42.9 (2016), pp. 2043–2049. DOI: 10.1016/j.ultrasmedbio.2016.04.001.

[211] Rosy Setiawati and Paulus Rahardjo. "Bone development and growth". In: *Osteogenesis and bone regeneration* 10 (2019). DOI: 10.5772/intechopen.82452.

[212] Rustam N. Karanjia, Mary M. E. Crossey, I. Jane Cox, Haddy K. S. Fye, Ramou Njie, Robert D. Goldin, and Simon D. Taylor-Robinson. "Hepatic steatosis and fibrosis: Non-invasive assessment". In: *World journal of gastroenterology* 22.45 (2016), p. 9880. DOI: 10.3748/wjg.v22.i45.9880.

[213] Zobair Younossi, Quentin M. Anstee, Milena Marietti, Timothy Hardy, Linda Henry, Mohammed Eslam, Jacob George, and Elisabetta Bugianesi. "Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention". In: *Nature reviews Gastroenterology & hepatology* 15.1 (2018), pp. 11–20. DOI: 10.1038/nrgastro.2017.109.

[214] Zachary D. Goodman. "Grading and staging systems for inflammation and fibrosis in chronic liver diseases". In: *Journal of hepatology* 47.4 (2007), pp. 598–607. DOI: 10.1016/j.jhep.2007.07.006.

[215] Qian Li, Manish Dhyani, Joseph R. Grajo, Claude Sirlin, and Anthony E. Samir. "Current status of imaging in nonalcoholic fatty liver disease". In: *World journal of hepatology* 10.8 (2018), p. 530. DOI: 10.4254/wjh.v10.i8.530.

[216] Yasushi Honda, Masato Yoneda, Kento Imajo, and Atsushi Nakajima. "Elastography techniques for the assessment of liver fibrosis in non-alcoholic fatty liver disease". In: *International journal of molecular sciences* 21.11 (2020), p. 4039. DOI: `10.3390/ijms21114039`.

[217] Golo Petzold, Julian Lasser, Janina Rühl, Sebastian C. B. Bremer, Richard F. Knoop, Volker Ellenrieder, Steffen Kunsch, and Albrecht Neesse. "Diagnostic accuracy of B-Mode ultrasound and Hepatorenal Index for graduation of hepatic steatosis in patients with chronic liver disease". In: *PloS one* 15.5 (2020), e0231044. DOI: `10.1371/journal.pone.0231044`.

[218] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. "Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications". In: *Medical physics* 47.5 (2020), e148–e167.

[219] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. "Perceiver: General perception with iterative attention". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4651–4664.

[220] Payal Dhar. "The carbon impact of artificial intelligence". In: *Nature Machine Intelligence* 2.8 (2020), pp. 423–425.

[221] Juliane R Sempionatto, Muyang Lin, Lu Yin, Kexin Pei, Thitaporn Sonsa-ard, Andre N de Loyola Silva, Ahmed A Khorshed, Fangyu Zhang, Nicholas Tostado, Sheng Xu, et al. "An epidermal patch for the simultaneous monitoring of haemodynamic and metabolic biomarkers". In: *Nature Biomedical Engineering* 5.7 (2021), pp. 737–748.

[222] Daniel Berger, Vishal Desai, and Sujit Janardhan. "Con: liver biopsy remains the gold standard to evaluate fibrosis in patients with nonalcoholic fatty liver disease". In: *Clinical liver disease* 13.4 (2019), p. 114.

[223] Shaham Mumtaz, Nathan Schomaker, and Natasha Von Roenn. "Pro: noninvasive imaging has replaced biopsy as the gold standard in the evaluation of nonalcoholic fatty liver disease". In: *Clinical Liver Disease* 13.4 (2019), p. 111.

[224] US Food, Drug Administration, et al. "Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan". In: *Published online December* 1 (2021).

[225] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1 (2021), p. 18.

[226] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf`.

[227] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. “Problems with Shapley-value-based explanations as feature importance measures”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5491–5500.

[228] Soley Hafthorsdottir, Sergei Vostrikov, Andrea Cossettini, Michael Rieder, Christoph Leitner, Michele Magno, and Luca Benini. “Automatic Extraction of Muscle Fascicle Pennation Angle from Raw Ultrasound Data”. In: *2022 IEEE Sensors Applications Symposium (SAS)*. IEEE. 2022, pp. 1–5.

[229] Victor R Lee. “The Quantified Self (QS) movement and some emerging opportunities for the educational technology field”. In: *Educational Technology* (2013), pp. 39–42.

[230] Ulysse Cote-Allard, Cheikh Latyr Fall, Alexandre Campeau-Lecours, Clément Gosselin, François Laviolette, and Benoit Gosselin. “Transfer learning for sEMG hand gestures recognition using convolutional neural networks”. In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. (2017), pp. 1663–1668. DOI: `10.1109/SMC.2017.8122854`.

[231] Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. “The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances”. In: *Data Mining and Knowledge Discovery* 35.2 (2021), pp. 401–449.

Appendix **A**

Appendix A

This appendix provides detailed results of muscle fatigue signal classifications in terms of accuracy in Section A.1 and in terms of speed for training and evaluation in Section A.2 as described in Chapter 3.

A.1. Accuracies for muscle fatigue state classifications

This section shows the accuracy results for several data type / ML model combinations based on a variety of different input signals.

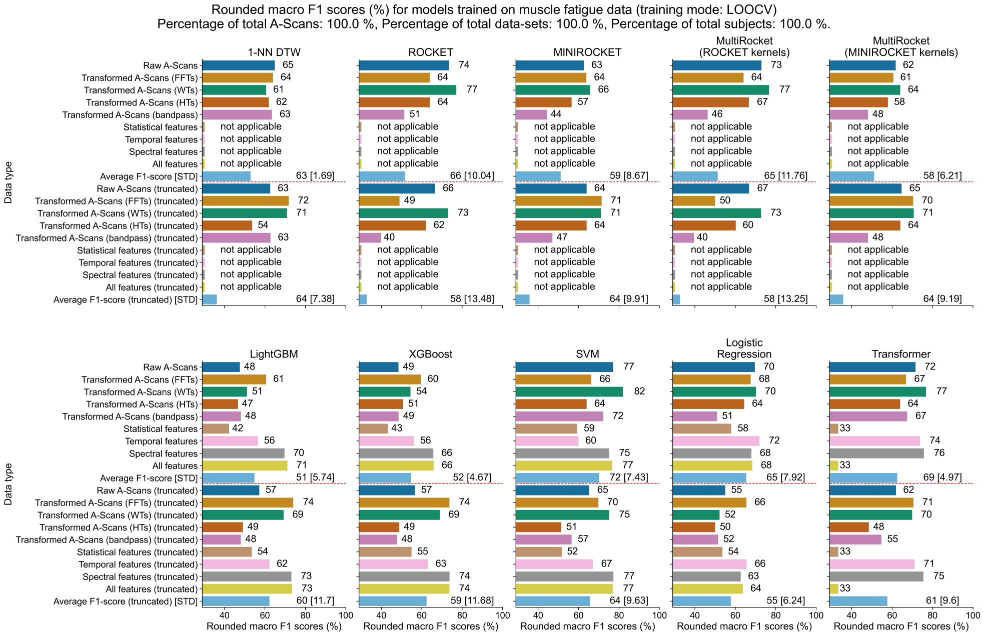


Figure A.1.: Accuracy results for all data type / ML model combinations of all signals.

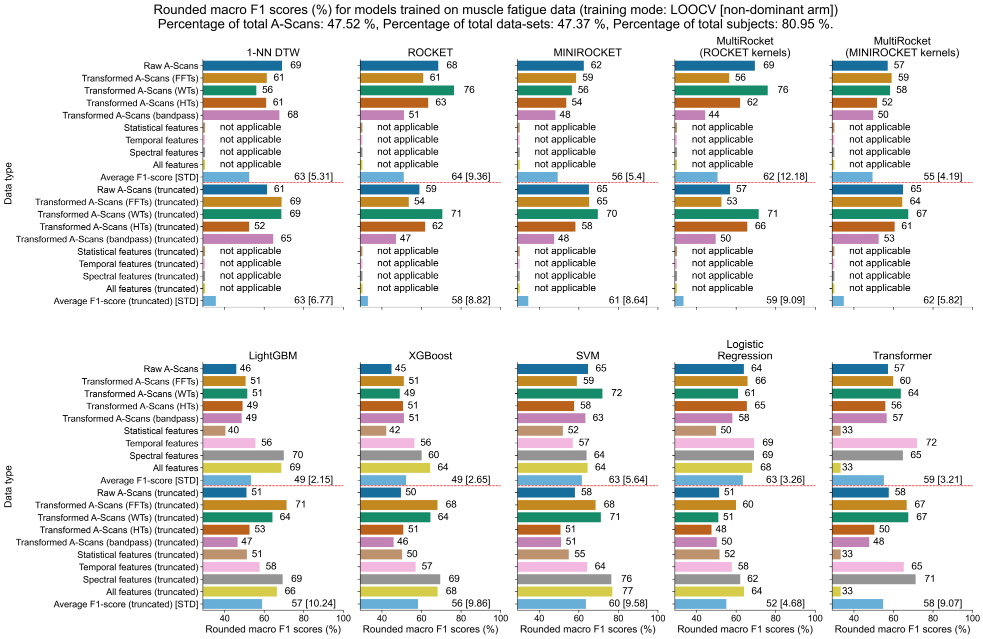


Figure A.2.: Accuracy results for all data type / ML model combinations of all signals from the non-dominant arm.

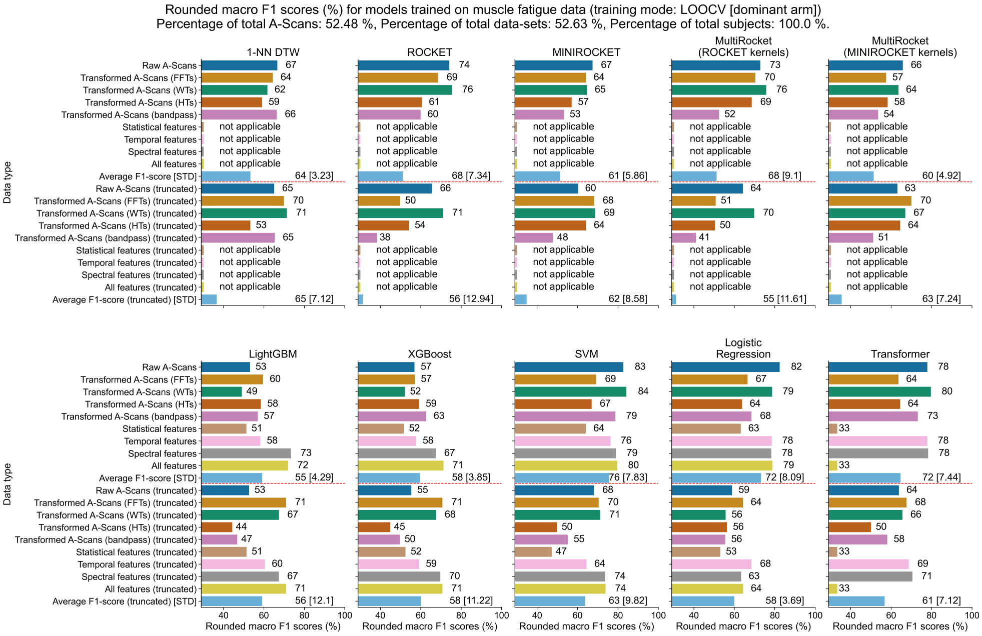


Figure A.3.: Accuracy results for all data type / ML model combinations of all signals from the dominant arm.

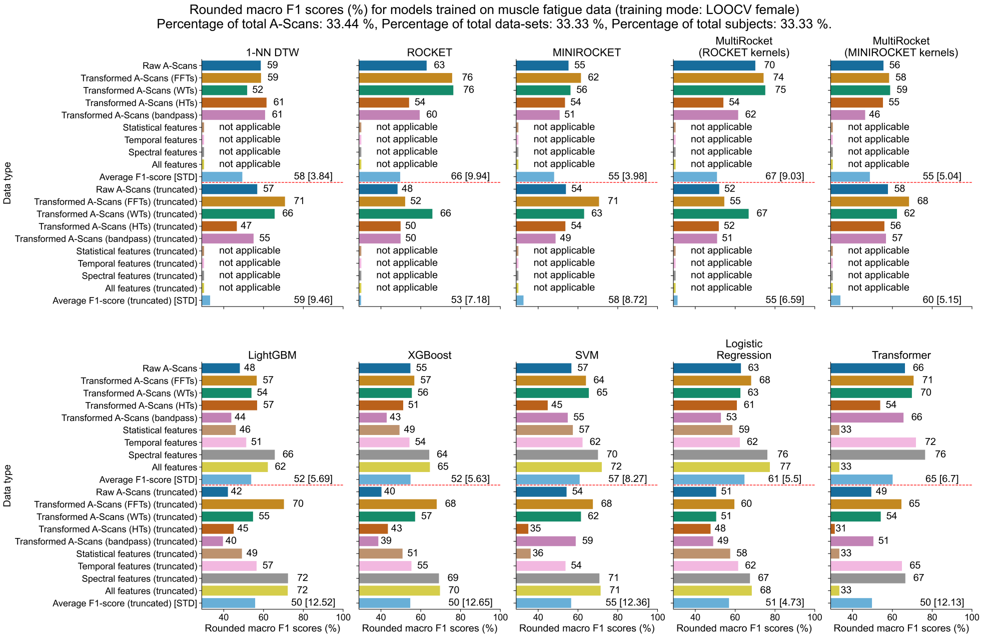


Figure A.4.: Accuracy results for all data type / ML model combinations of all signals from female subjects only.

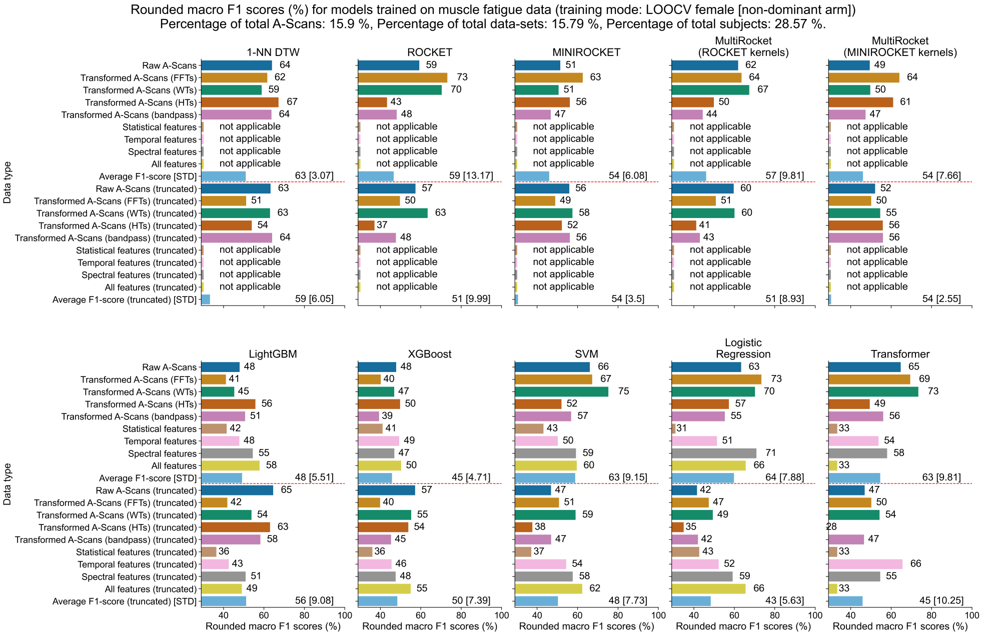


Figure A.5.: Accuracy results for all data type / ML model combinations of all signals from the non-dominant arm of female subjects only.

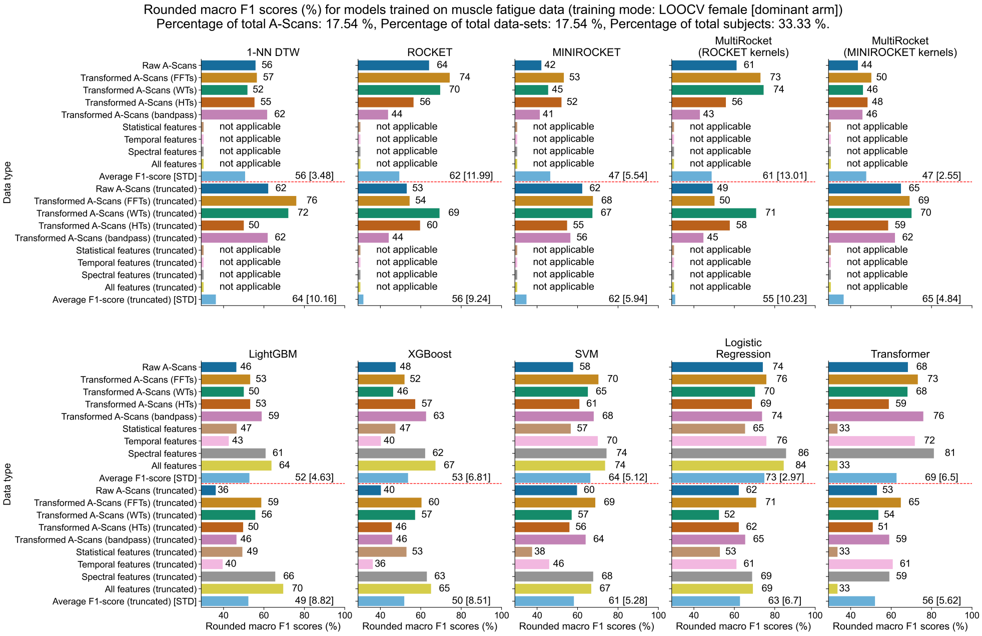


Figure A.6.: Accuracy results for all data type / ML model combinations of all signals from the dominant arm of female subjects only.

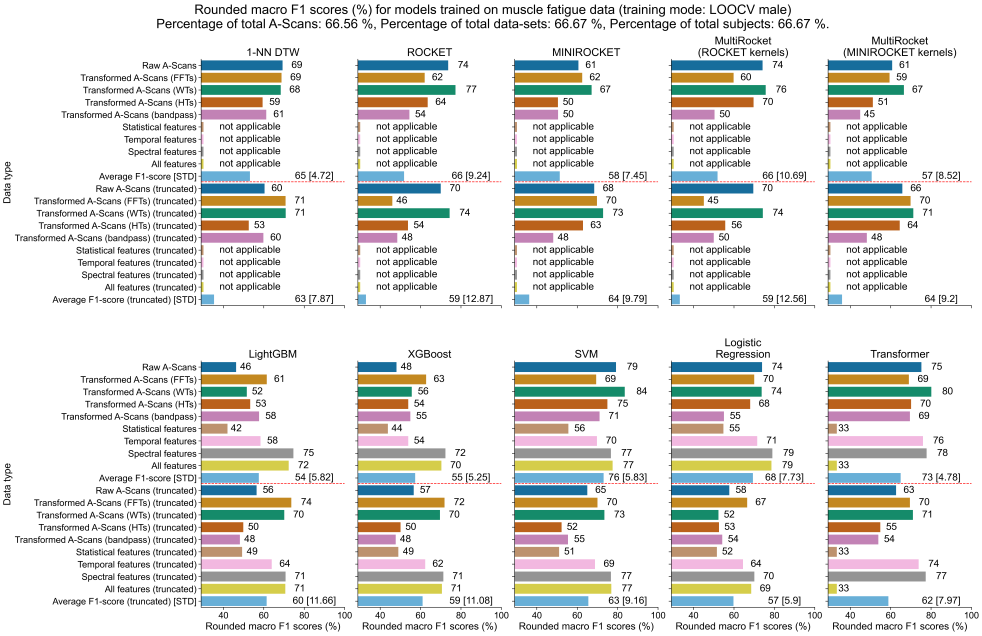


Figure A.7.: Accuracy results for all data type / ML model combinations of all signals from male subjects only.

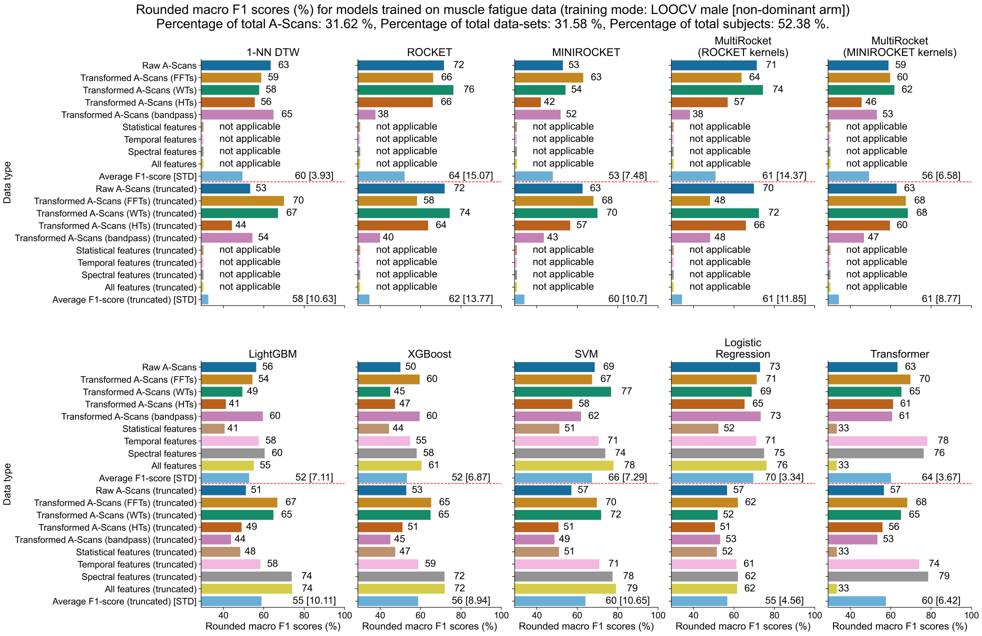


Figure A.8.: Accuracy results for all data type / ML model combinations of all signals from the non-dominant arm of male subjects only.

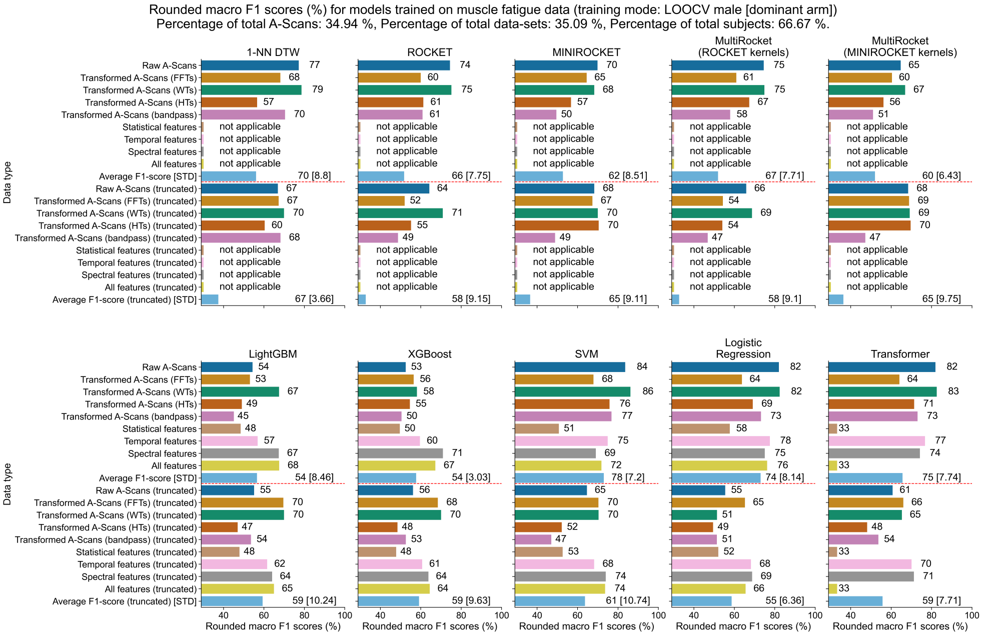


Figure A.9.: Accuracy results for all data type / ML model combinations of all signals from the dominant arm of male subjects only.

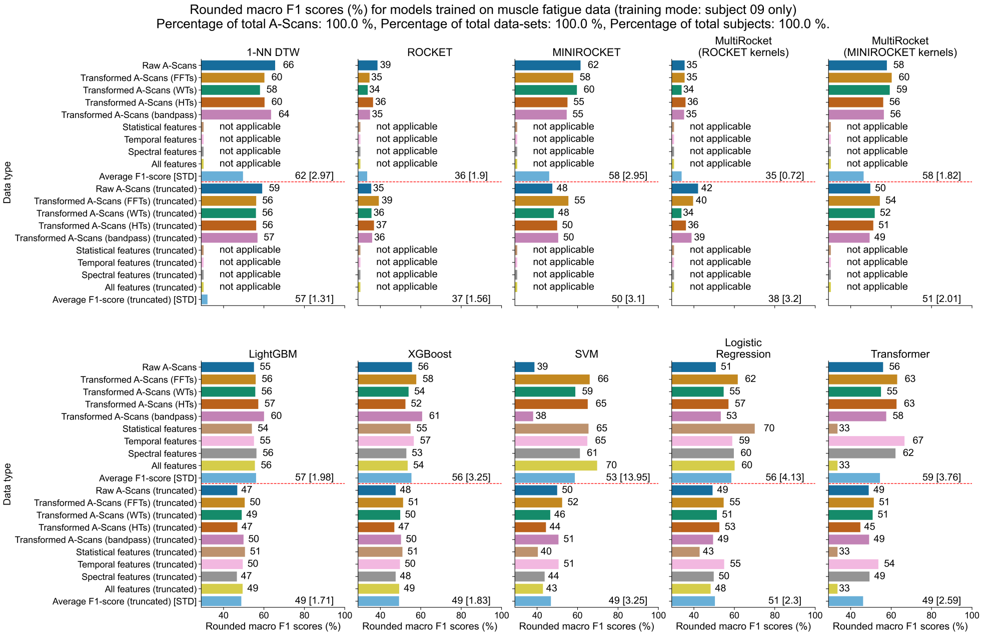


Figure A.10.: Accuracy results for all data type / ML model combinations of all signals stemming from a single subject only.

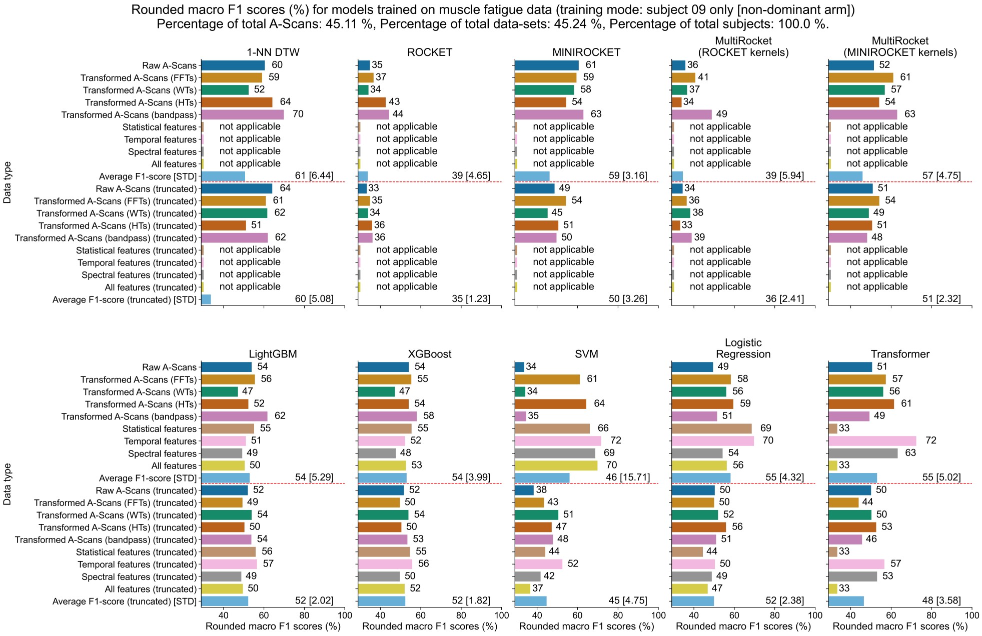


Figure A.11.: Accuracy results for all data type / ML model combinations of all signals stemming from the non-dominant arm of a single subject only.

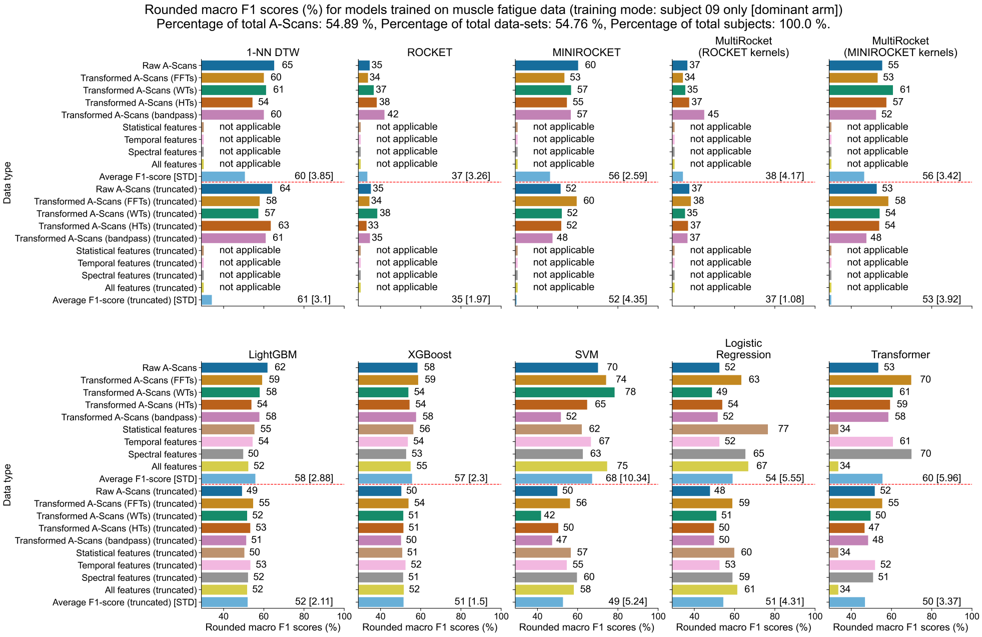


Figure A.12.: Accuracy results for all data type / ML model combinations of all signals stemming from the dominant arm of a single subject only.

A.2. Evaluation and training times for muscle fatigue state classifications

This section shows the time for training and evaluation for several data type / ML model combinations based on a variety of different input signals.

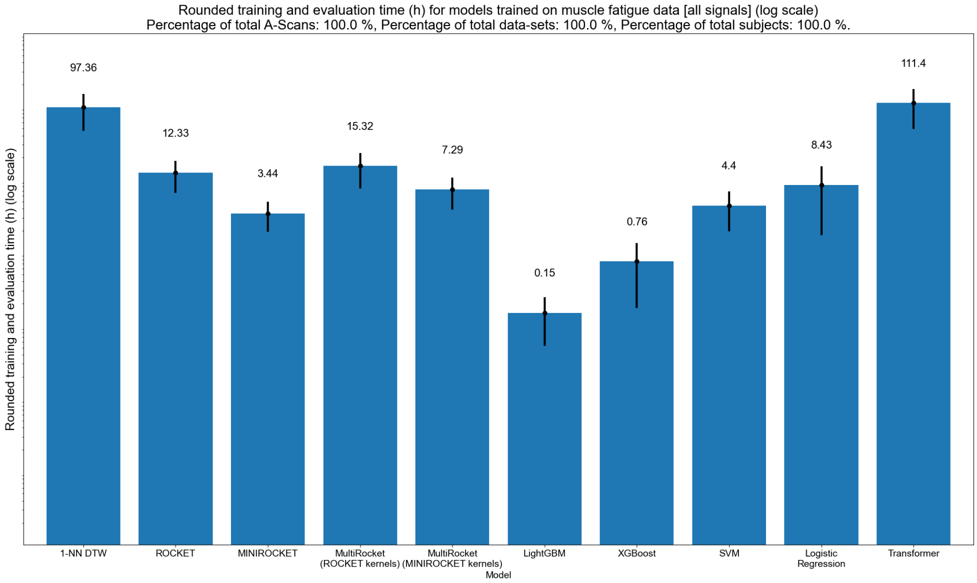


Figure A.13.: Time for training and evaluation for all data type / ML model combinations of all signals.

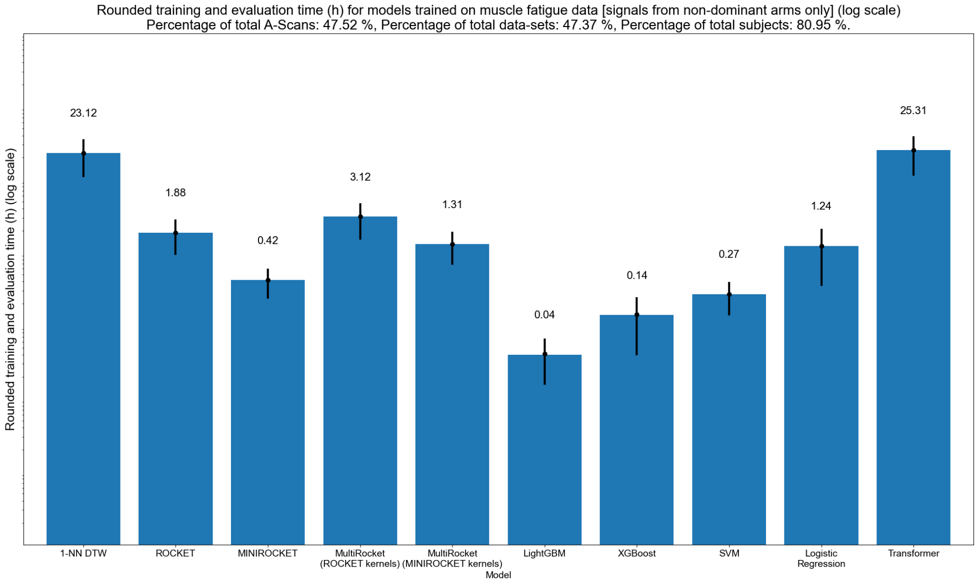


Figure A.14.: Time for training and evaluation for all data type / ML model combinations of all signals from the non-dominant arm.

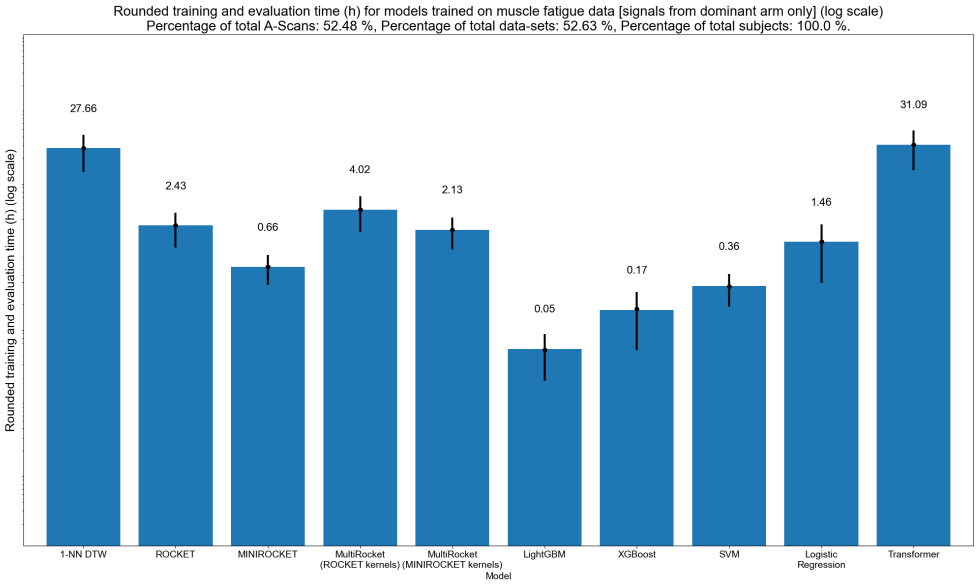


Figure A.15.: Time for training and evaluation for all data type / ML model combinations of all signals from the dominant arm.

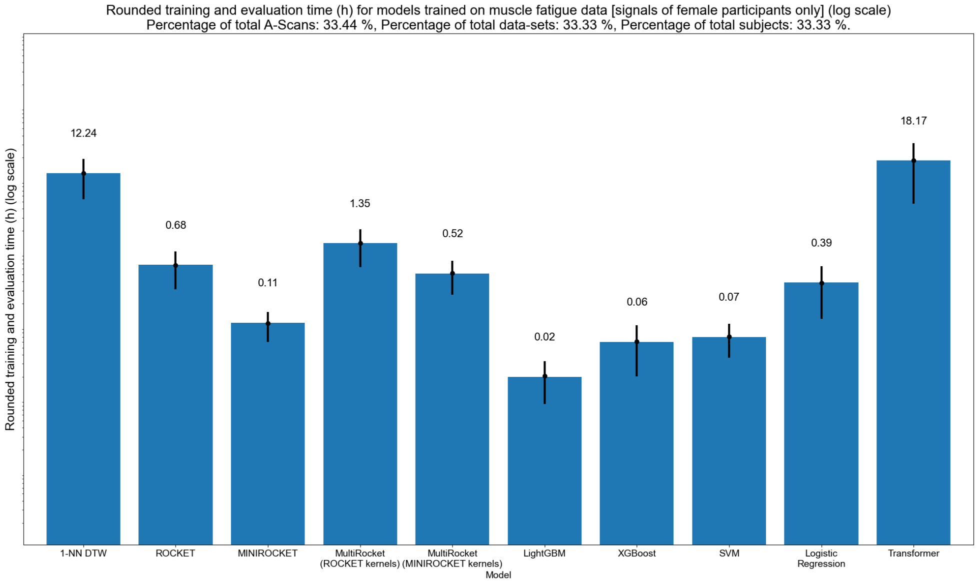


Figure A.16.: Time for training and evaluation for all data type / ML model combinations of all signals from female subjects only.

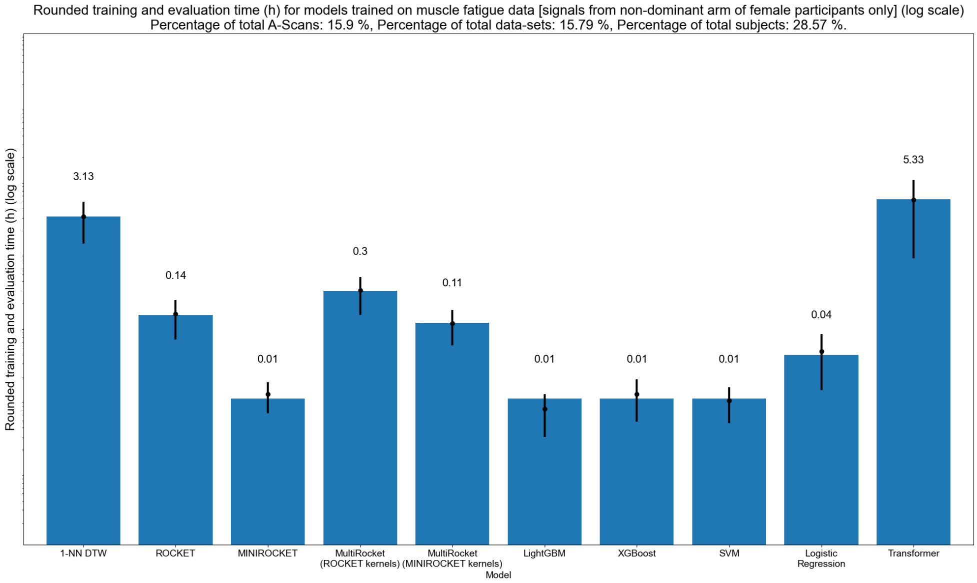


Figure A.17.: Time for training and evaluation for all data type / ML model combinations of all signals from the non-dominant arm of female subjects only.

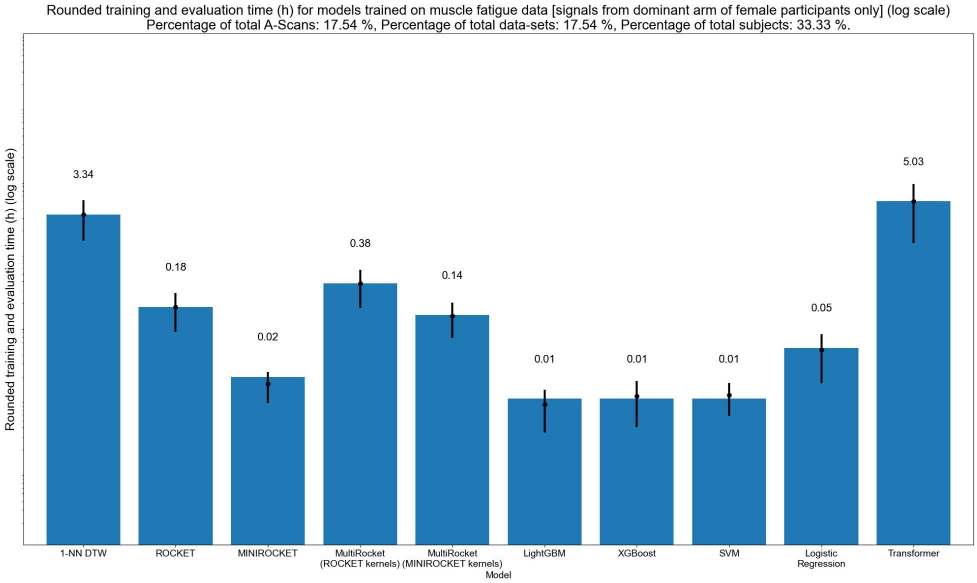


Figure A.18.: Time for training and evaluation for all data type / ML model combinations of all signals from the dominant arm of female subjects only.

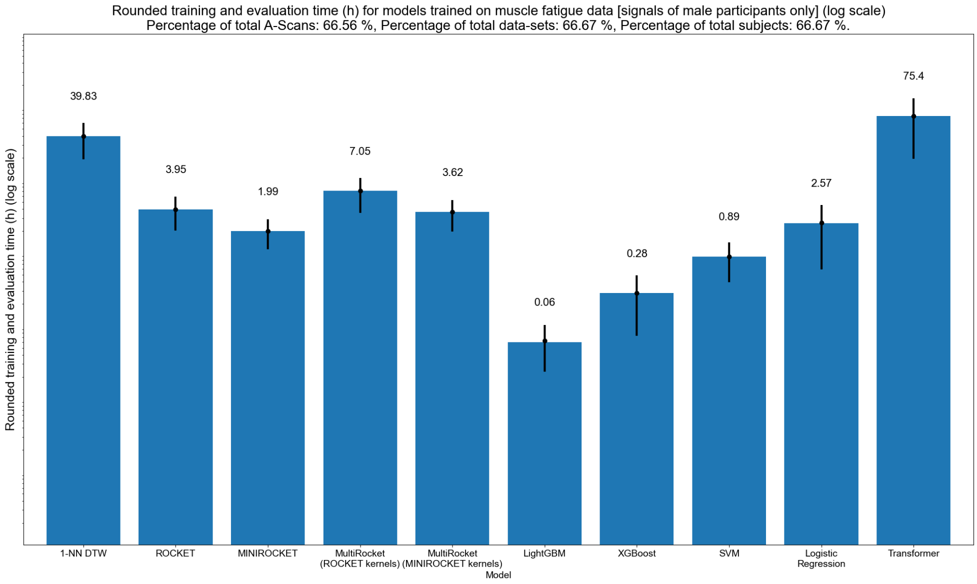


Figure A.19.: Time for training and evaluation for all data type / ML model combinations of all signals from male subjects only.

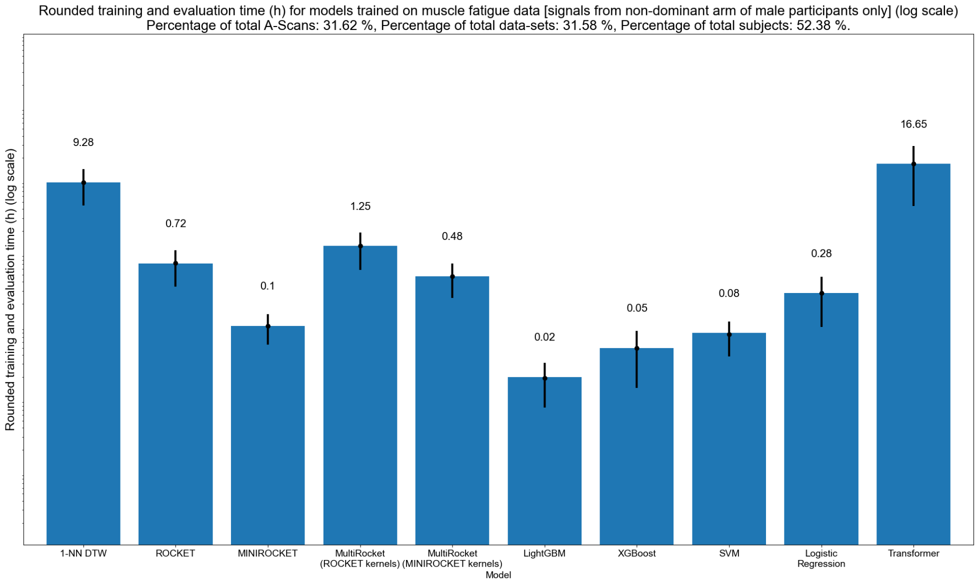


Figure A.20.: Time for training and evaluation for all data type / ML model combinations of all signals from the non-dominant arm of male subjects only.

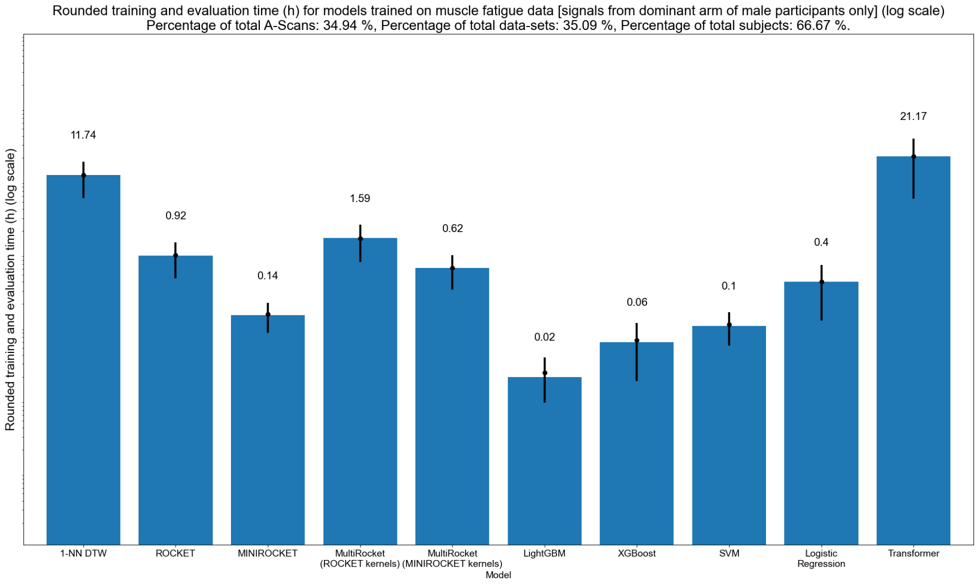


Figure A.21.: Time for training and evaluation for all data type / ML model combinations of all signals from the dominant arm of male subjects only.

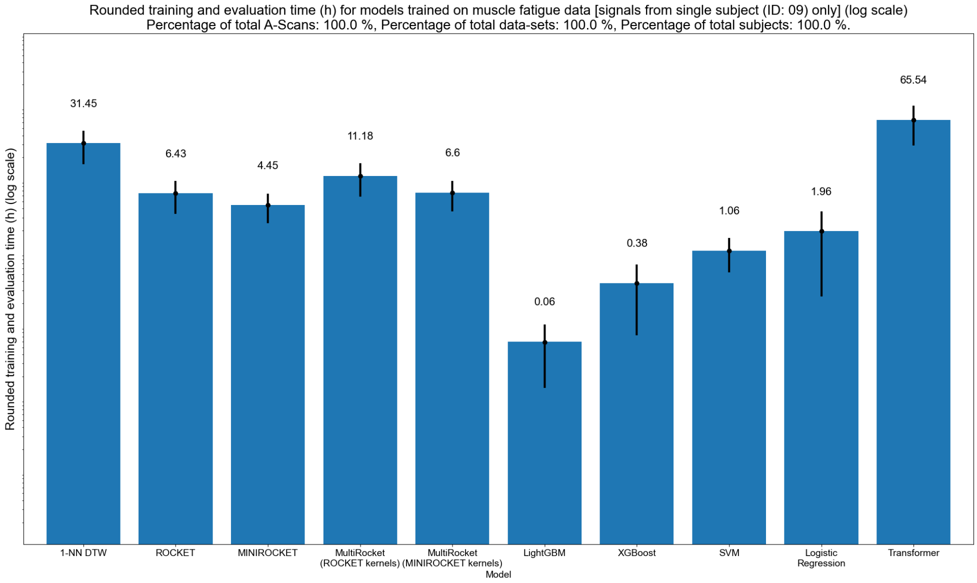


Figure A.22.: Time for training and evaluation for all data type / ML model combinations of all signals stemming from a single subject only.

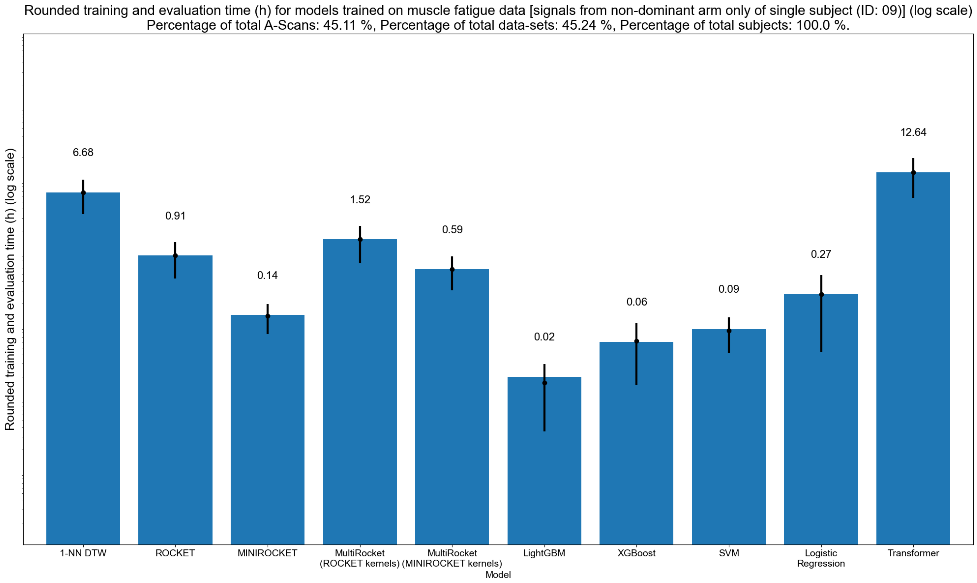


Figure A.23.: Time for training and evaluation for all data type / ML model combinations of all signals stemming from the non-dominant arm of a single subject only.

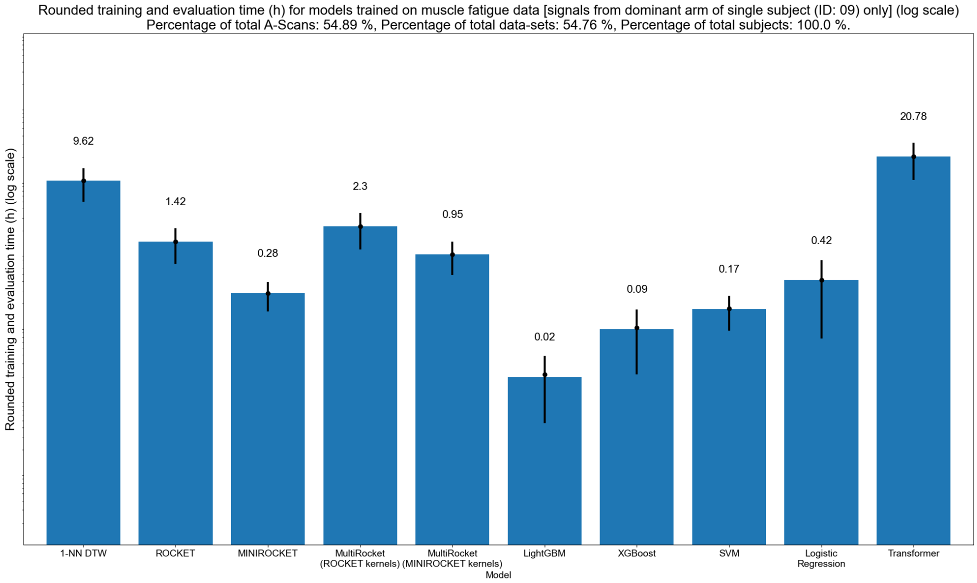


Figure A.24.: Time for training and evaluation for all data type / ML model combinations of all signals stemming from the dominant arm of a single subject only.

Appendix B

Results for detection of epiphyseal radius bone closure

This appendix provides detailed results of epiphyseal radius bone closure classifications based on acquired US as described in Chapter 4. To this end, Table B.1 presents the F_1-score for each ML model and each study subject. F_1 scores $\geq$ 80% are colored in green, F_1 scores between 50% and 80% are colored in yellow and all other cells are colored in red.

ID	Age in years	1-NN DTW	RBF NN	MLP	FCN	ResNet	XGBoost	Light GBM	Cat Boost with 100 iter.	Cat Boost with 1,000 iter.	Cat Boost with 10,000 iter.	SVM	LR
1	9.25	100	88	100	100	100	84	76	100	100	100	100	100
2	9.50	100	100	100	100	100	100	100	100	100	100	100	4
3	9.60	100	100	0	100	100	52	80	40	24	72	100	0
4	10.25	100	100	100	100	100	100	84	100	100	100	100	100
5	10.50	48	88	16	36	100	100	100	100	100	100	68	40
6	11.00	100	4	100	100	0	44	56	64	92	96	20	4
7	11.00	100	100	100	100	0	96	40	100	100	100	100	40
8	11.00	100	20	84	56	64	68	96	52	100	100	68	24
9	12.00	100	100	100	100	0	100	48	100	100	100	100	100
10	12.25	100	100	100	100	100	100	100	100	100	100	100	100
11	12.50	48	0	0	100	100	100	100	96	100	100	0	44
12	12.50	100	100	76	100	100	100	96	100	100	100	100	80
13	12.60	0	100	100	100	96	92	88	100	100	100	100	100
14	12.75	100	76	84	72	80	100	92	100	96	100	100	64
15	13.00	96	100	100	100	0	96	92	100	100	100	100	100
16	13.25	100	100	100	100	100	100	100	100	100	100	100	100
17	13.25	100	100	88	100	100	100	100	100	92	96	60	0
18	13.60	56	44	100	100	100	100	100	100	100	100	100	100
19	13.75	60	0	100	100	100	100	100	100	100	100	80	56
20	14.00	28	0	100	80	92	92	100	72	68	68	52	80
21	14.00	44	24	0	100	100	100	100	100	100	100	100	20
22	14.90	0	0	0	0	0	20	64	0	0	0	0	0
23	15.00	60	100	100	100	0	8	0	76	28	32	100	84
24	15.00	36	16	100	100	100	48	0	44	16	12	60	100
25	15.00	32	60	100	100	0	100	100	100	100	100	100	100
26	15.00	56	60	100	100	0	100	100	100	100	100	36	40
27	15.00	0	0	0	96	100	100	100	24	92	88	0	76
28	15.00	68	0	84	100	100	100	100	100	100	100	100	100
29	15.10	0	8	0	0	4	4	12	12	4	4	4	4
30	15.50	72	0	24	72	96	100	0	100	100	100	92	100
31	15.60	100	0	36	0	0	52	68	12	0	0	0	0
32	15.60	8	80	64	100	100	52	44	84	88	84	100	100
33	16.00	44	56	100	60	48	100	100	100	100	100	100	100
34	16.00	80	0	0	0	8	100	100	100	100	100	72	84
35	16.00	100	0	12	100	100	68	88	76	60	64	0	0
36	16.00	48	16	36	0	0	92	40	84	80	84	0	32
37	16.60	44	20	56	0	0	0	0	92	64	96	100	100
38	17.00	76	100	100	56	100	100	100	100	100	100	100	84
39	17.00	36	0	0	8	8	52	100	0	0	0	0	0

ID	Age in years	1-NN DTW	RBF NN	MLP	FCN	ResNet	XGBoost	Light GBM	Cat Boost with 100 iter.	Cat Boost with 1,000 iter.	Cat Boost with 10,000 iter.	SVM	LR
40	17.30	96	28	24	0	0	0	0	16	4	0	0	48
41	17.50	88	0	96	100	100	100	100	100	100	100	100	100
42	17.70	0	0	0	0	0	0	0	0	0	0	0	0
43	18.00	44	60	44	40	48	4	8	36	60	76	60	32
44	18.00	96	52	100	100	100	100	100	100	100	100	56	100
45	18.10	100	100	80	4	0	56	8	100	100	100	20	0
46	18.20	100	100	100	100	100	100	100	100	100	100	100	100
47	18.30	100	100	100	100	100	100	100	100	100	100	12	68
48	18.60	100	100	80	100	100	96	12	92	100	100	100	100
49	18.60	0	0	0	0	0	100	100	100	100	100	0	0
50	18.70	100	100	100	100	100	100	100	100	100	100	100	100
51	18.80	100	100	100	100	88	100	80	100	100	100	100	100
52	19.00	100	100	60	28	44	20	60	56	56	56	16	0
53	19.00	64	100	72	0	0	100	100	4	96	52	0	100
54	19.10	100	100	100	100	100	100	100	100	100	100	100	100
55	19.20	88	96	48	100	100	100	100	100	100	100	100	100
56	19.20	100	100	100	100	100	100	100	100	100	100	100	100
57	19.50	48	100	100	24	92	8	8	24	32	28	72	32
58	19.60	100	88	76	100	100	100	100	100	100	100	52	28
59	19.60	100	100	100	100	100	100	100	100	100	100	100	100
60	19.60	88	96	96	100	100	100	100	100	100	100	100	100
61	19.60	100	100	100	100	100	100	100	100	100	100	100	100
62	19.60	100	84	100	100	100	100	100	100	100	100	100	96
63	19.70	100	100	100	100	100	100	100	100	100	100	100	100
64	19.80	100	100	76	100	100	100	100	100	100	100	100	100
65	19.90	100	100	100	100	100	72	100	96	100	100	100	100
66	20.00	48	56	96	52	76	32	16	32	28	32	60	8
67	20.00	60	100	36	96	84	20	60	16	24	32	84	32
68	20.00	100	100	100	100	100	100	100	100	100	100	100	100
69	20.00	100	100	100	100	100	100	96	100	100	100	100	100
70	20.00	100	100	88	100	100	100	100	100	100	100	100	100
71	20.00	100	100	100	100	100	100	100	100	100	100	100	72
72	20.00	100	100	60	100	92	100	100	100	100	100	100	40
73	20.10	100	100	100	100	100	100	100	100	100	100	100	100
74	20.10	100	92	100	100	96	100	100	100	100	100	100	100
75	20.20	100	100	92	100	100	100	100	100	100	100	100	100
76	20.20	100	100	100	100	100	100	100	100	100	100	100	100
77	20.20	100	100	100	84	100	100	32	100	100	100	100	100
78	20.30	100	100	72	52	64	100	100	100	100	100	96	76
79	20.30	100	100	100	100	100	100	100	100	100	100	100	100
80	20.30	100	100	100	100	96	100	100	100	100	100	100	100
81	20.50	100	100	100	100	100	100	100	100	100	100	100	100
82	20.60	100	100	100	100	100	100	100	100	100	100	100	100
83	20.60	100	100	100	100	100	100	100	100	100	100	100	64
84	20.60	100	100	100	100	100	100	100	100	100	100	100	100
85	20.70	100	100	100	100	100	100	100	100	100	100	100	100
86	20.70	100	100	100	100	100	100	100	100	100	100	100	100
87	20.70	100	100	100	100	100	100	100	100	100	100	100	80
88	20.70	100	100	100	100	100	100	100	100	100	100	100	100
89	20.70	100	100	100	100	100	100	100	100	100	100	100	100
90	20.70	100	68	60	100	100	100	100	100	100	100	60	72
91	20.80	100	100	100	100	100	100	100	100	100	100	100	100
92	20.80	100	0	40	100	100	100	100	100	100	100	100	96
93	20.80	100	100	100	100	100	100	100	100	100	100	100	100
94	20.90	100	100	100	100	100	100	100	100	100	100	100	100
95	20.90	52	100	100	100	100	100	100	100	100	100	100	100
96	20.90	88	100	100	100	100	100	100	100	100	100	100	100
97	20.90	100	64	100	100	100	100	100	100	100	100	96	100
98	20.90	100	0	0	100	100	100	100	100	100	100	0	100
99	21.00	100	100	100	0	52	48	16	32	48	52	52	68
100	21.10	100	100	100	100	100	100	100	100	100	100	100	100
101	21.20	100	72	100	96	12	92	92	12	36	68	100	100
102	21.30	100	100	100	100	100	100	100	100	100	100	100	100
103	21.30	100	100	100	100	100	100	100	100	100	100	100	100
104	21.30	100	100	100	100	80	100	96	100	100	100	100	100
105	21.30	68	0	0	4	0	0	0	0	0	0	0	0
106	21.50	16	68	0	36	8	0	0	24	68	36	28	92
107	21.60	100	88	100	100	100	100	100	100	100	100	100	100
108	21.70	100	100	80	100	100	100	96	100	100	100	100	100
109	21.70	100	100	100	100	100	100	100	100	100	100	100	100
110	21.70	4	100	0	24	96	60	92	4	12	4	0	8
111	22.00	100	100	100	100	100	100	100	100	100	100	88	68
112	22.10	100	100	100	100	100	100	100	100	100	100	100	100
113	22.10	100	24	100	100	100	100	100	100	100	100	52	48
114	22.40	100	56	100	100	100	100	100	100	100	100	76	100
115	22.60	100	100	100	100	100	100	100	100	100	100	100	100
116	22.60	76	20	100	100	100	100	76	100	100	100	88	100
117	22.70	20	0	8	100	100	100	100	100	100	100	24	36
118	22.80	100	100	84	100	100	100	96	100	100	100	100	100
119	22.90	100	100	100	100	100	100	100	100	100	100	80	100

ID	Age in years	1-NN DTW	RBF NN	MLP	FCN	ResNet	XGBoost	Light GBM	Cat Boost with 100 iter.	Cat Boost with 1,000 iter.	Cat Boost with 10,000 iter.	SVM	LR
120	23.90	96	100	100	100	100	100	100	100	100	100	100	100

Table B.1.: F_1-scores for each ML model and each study subject.

Appendix C

Clinical study to identify hepatic steatosis and fibrosis in patients with non-alcoholic fatty liver disease

This appendix provides a comprehensive overview of all data collected for each participant of the clinical study to identify hepatic steatosis and fibrosis in patients with non-alcoholic fatty liver disease (see Chapter 5). Table C.1 provides an overview of the collected anamnesis data for each patient, while Table C.2 provides detailed liver values for each patient. Table C.3 shows liver health assessments for each participant as provided by medical practitioners for the clinical study. Table C.4 provides acquired *FibroScan* and ARFI values for each patient.

Subject ID	**Weight (kg)**	**Height (m)**	**BMI (kg/m^2)**	**Hip girth (cm)**	**Waist circumference (cm)**	**RR (mmHg)**	**Heart frequency (/min)**	**Diabetes (Y=1, N=0)**
1	111.00	1.82	33.51	114	116	136/80	61	1
2	99.50	1.68	35.25	129	104	120/78	76	0
3	105.00	1.72	35.49	123	111	147/85	82	0
4	88.00	1.62	33.53	116	100	134/97	72	0
5	76.00	1.66	27.58	89	93	118/72	66	1
6	85.00	1.74	28.08	100	99	140/95	113	0
7	112.00	1.83	33.44	116	113	134/82	71	1
8	94.00	1.65	34.53	125	114	133/76	67	0
9	78.00	1.64	29.00	106	114	148/93	74	1
10	73.00	1.65	26.81	98	92	112/71	83	1
11	81.30	1.76	26.25	103	91	121/73	52	1
12	91.00	1.70	31.49	106	104	145/93	55	0
13	103.00	1.84	30.42	101	108	135/90	71	1
14	90.00	1.75	29.39	110	105	115/79	73	1
15	87.00	1.75	28.41	98	102	118/66	57	0
16	81.00	1.86	23.41	95	95	118/78	69	0
17	116.00	1.77	37.03	108	110	135/85	72	0
18	102.00	1.80	31.48	105	113	117/83	54	1
19	112.00	1.85	32.72	106	108	162/88	56	0
20	112.00	1.75	36.57	121	112	133/82	130	0
21	86.00	1.75	28.08	91	101	140/92	56	1
22	77.00	1.66	27.94	94	90	111/63	51	0
23	93.00	1.70	32.18			138/81	99	1
24	115.00	1.88	32.54	105	107	119/76	72	0
25	116.00	1.81	35.41					0
26	113.00	1.75	36.90			126/69	57	0
27	87.00	1.72	29.41					1

Table C.1.: Anamnesis data

Subject ID	GOT (AST)	GPT (ALT)	GOT/GPT	gGT	Total bilirubin	Thrombocytes	Albumin (g/dl)
1	33	46	0.717391304	68	0.4	254	5.2
2	52	39	1.333333333	21	0.6	189	4.6
3	88	124	0.709677419	780	0.3	333	5
4	22	21	1.047619048	15	0.2	217	4.7
5	45	44	1.022727273	255	0.8	146	4.6
6	46	89	0.516853933	57	0.6	265	5
7	59	57	1.035087719	64	0.5	144	4.6
8	25	41	0.609756098	368	0.4	189	4.3
9	36	43	0.837209302	93	0.7	193	4.4
10	15	16	0.9375	52	0.2	221	4.1
11	61	140	0.435714286	44	0.5	247	5.2
12	43	53	0.811320755	31	0.7	208	4.9
13	47	136	0.345588235	53	0.7	226	5
14	45	55	0.818181818	177	0.5	358	4.5
15	31	53	0.58490566	31	0.5	134	5
16	37	34	1.088235294	214	0.3	227	4.1
17	67	67	1	308	0.3	225	4.5
18	172	158	1.088607595	203	1	104	4.8
19	48	52	0.923076923	156	1.3	143	4.3
20	30	44	0.681818182	36	0.4	348	4.4
21	33	83	0.397590361	285	0.6	215	4.9
22	42	44	0.954545455	51	0.4	244	4.8
23	49	45	1.088888889	54	0.8	148	3.6
24	44	75	0.586666667	16	2.2	232	5
25	45	47	0.957446809	43	0.7	253	4.7
26	33	99	0.333333333	255	1.4	332	4
27	47	90	0.522222222	28	0.3	245	4.2

Table C.2.: Liver values

Subject ID	NFS	Fibrosis stage assessment	BARD	Risk assessment	FIB-4	Fibrosis assessment
1	-2.809800722	F0-2	2	high risk	0.900324701	no advanced fibrosis
2	-0.610153628	indeterminate	3	high risk	2.29093401	indeterminate
3	-3.23015651	F0-2	1	Low risk	1.305239732	no advanced fibrosis
4	-1.3368986	indeterminate	3	high risk	1.238914474	no advanced fibrosis
5	-0.41396117	indeterminate	3	high risk	3.252605258	probable advanced fibrosis
6	-3.419260571	F0-2	2	high risk	0.91999816	no advanced fibrosis
7	-0.305544325	indeterminate	4	high risk	3.093334942	indeterminate
8		indeterminate	1	Low risk	1.260132206	no advanced fibrosis
9		indeterminate	4	high risk	1.991173043	indeterminate
10		F0-2	3	high risk	0.865384615	no advanced fibrosis
11		F0-2	0	Low risk	0.459189073	no advanced fibrosis
12		F0-2	3	high risk	1.533419373	indeterminate
13	-3.712108668	F0-2	1	Low risk	0.481486072	no advanced fibrosis
14	-3.46955102	F0-2	3	high risk	1.033898672	no advanced fibrosis
15	-1.469576049	F0-2	1	low risk	1.715981971	indeterminate
16	-2.462814593	F0-2	2	high risk	1.202001097	no advanced fibrosis
17	-1.212518657	indeterminate	4	high risk	1.855346628	indeterminate
18	0.024980778	indeterminate	3	high risk	7.762803192	probable advanced fibrosis
19	0.133960106	indeterminate	4	high risk	3.165282239	indeterminate
20	-3.251285714	F0-2	1	low risk	0.610820396	no advanced fibrosis
21	-2.820712073	F0-2	1	low risk	0.842376879	no advanced fibrosis
22	-2.075348817	F0-2	1	low risk	1.6607838	indeterminate
23					2.86	
24	-3.389689362	F0-F2			0.569386817	no advanced fibrosis
25	-1.532779929	indeterminate	4	high risk	1.582606331	indeterminate
26	-3.241591837	F0-F2	1	Low risk	0.429562849	no advanced fibrosis
27	-2.759667929	F0-F2	1	Low risk	0.869518796	no advanced fibrosis

Table C.3.: Liver health assessment by medical practitioners

Subject ID	**Fibroscan probe type**	**Fibroscan CAP median (dB/m)**	**Fibroscan min CAP**	**Fibroscan max CAP**	**Fibroscan E (kPa)**	**ARFI (m/s)**	**ARFI (SD)**
1	XL	362	304	392	5.8	0.84	0.15
2	XL	272	191	400	7.9	1.14	0.07
3	XL	318	248	394	18.8	2.47	0.43
4	XL	326	257	354	3.3	1.15	0.25
5	XL	313	278	368	48.8	2.83	0.71
6	M	377	345	400	7.8	1.32	0.12
7	M	400	332	400	21.3	1.46	0.21
8	XL	347	320	389	12.2	0.63	0.09
9	M	340	306	373	11.6	1.01	0.09
10	XL	282	233	310	6.3	1.08	0.01
11	M	327	284	355	7.9	1.34	0.06
12	XL	236	181	295	9	1.37	0.24
13	M	315	293	375	6.8	1.15	0.08
14	M	263	214	294	5.3	0.8	0.06
15	M	231	207	258	5.3	0.96	0.15
16	M	208	177	255	6.3	1.21	0.09
17	XL	400	364	400	39	1.97	0.48
18	M	321	271	374	12.4	1.07	0.19
19	M	340	250	355	15.7	1.87	0.43
20	XL	383	308	400	3.8	1	0.27
21	M	285	238	318	4.5	0.78	0.12
22	M	299	280	320	6.2	1.24	0.11
23	M	293	263	336	31	2.7	0.59
24	XL	339	317	370	5.6	0.99	0.07
25	M	377	340	400	19.2	1.87	0.43
26						1.1	0.14
27						1.14	0.07

Table C.4.: *FibroScan* and ARFI values

Appendix D

Curriculum Vita

Work experience

Since 04/2017	**PhD student and research associate** Fraunhofer-Institut für Biomedizinische Technik IBMT, St.Ingbert Department of Ultrasound, Software Development / System Integration
2016-2017	**Master's student and auxiliary scientist** Fraunhofer-Institut für Biomedizinische Technik IBMT, St.Ingbert Department of Ultrasound, Software Development / System Integration
2012-2015	**Auxiliary scientist** Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken
2010-2012	**Auxiliary scientist at the computer graphics chair of Prof.Slusallek** University of Saarland, Saarbrücken
2009-2010	**Auxiliary scientist at the computer graphics chair of Prof.Hermanns** University of Saarland, Saarbrücken
2008-2009	**Computer teacher** Pandahill Secondary School, Songwe, Tanzania

Education

08/2016	**Master's degree Visual Computing** University of Saarland, Saarbrücken
04/2014	**Bachelor's degree Computer Science** University of Saarland, Saarbrücken
2008	**Abitur** Technical high school, Völklingen