INAUGURAL – DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht – Karls – Universität

Heidelberg

vorgelegt von

Große Sundrup, Jonas Christopher, M.Sc.

aus Münster

Tag der mündlichen Prüfung : ……

Classification of Human Motion with Applications in Gesture Recognition and Treatment of Major Depressive Disorder

Betreuer : Prof. Dr. Katja Mombaur
Prof. Dr. Knut Schnell

Contents

List of Figures

List of Tables

Zusammenfassung

In dieser Arbeit betrachten wir daher die Frage der Klassifikation menschlicher Bewegung aus drei Perspektiven: Zunächst entwickeln wir eine Methode, die es erlaubt, menschliche Muskelsignale unter Berücksichtigung natürlichen Zeitversatzes zur Unterscheidung der ihnen zugrunde liegenden Gesten miteinander zu vergleichen. Hierbei können wir zeigen, dass die unter Zuhilfenahme von Dynamic Time Warping entwickelte Methode zum einen in der Lage ist, Klassifikationsergebnisse gemäß dem Stand der Forschung zu liefern und zu übertreffen und darüberhinaus auch erlaubt, Einblicke in die zugrundeliegenden Mechanismen zu gewinnen, die einer solchen Klassifikationsleistung zugrunde liegen. Wir können weiterhin zeigen, wie sich basierend auf diesem Einblick intelligente Verbesserungen und Modifikationen des zugrundeliegenden Datensatzes ableiten lassen. Im nächsten Teil der Arbeit wenden wir uns der Klassifikation von indirekten Signalen zu, indem die Unterscheidbarkeit zwischen Bewegungsmustern innerhalb und außerhalb von akuten Stresszuständen in Form von Grübeln bei Depressionspatienten untersucht wird. Wir können hierbei zum Einen zeigen, dass eine automatisierte Optimierung eines problemspezifischen Klassifikationsverfahren erfolgreich durchführbar ist. Desweiteren können wir zeigen, dass eine entsprechende Auftrennung der Teilkomponenten eines Klassifikationsverfahren ebenfalls einen Einblick in die Funktionsweise des Verfahrens sowie eine informierte Auswahl von optimierten Teilkomponenten ermöglicht. Wir können in der Konsequenz ein leichtgewichtiges optimiertes Klassifikationsverfahren vorschlagen, dass die spezifische Struktur der erfassten Patientendaten berücksichtigt und eine für medizinische Anwendungen geeignete, weil verstehbare Klassifikation liefern kann. Schlussendlich untersuchen wir die Frage der optimalen Sensorlokation, um für Anwendungen im Zusammenhang mit Sensormessungen das entsprechende Sensorlayout unter gegebenen Limitationen optimal auszugestalten. Wir schlagen am Beispiel eines Rekonstruktionsproblems ein Verfahren vor, dass automatisiert die Qualität eines bestimmten Sensorlayouts für vergleichbare Anwendungen bewerten kann. Dies bildet dann die Grundlage für ein Verfahren, das autonom in der Lage ist, optimale Sensorlokationen für das zugrunde liegende Problem zu finden. Desweiteren können wir zeigen, dass die vorgeschlagene Methode sich eignet, um die Sensitivität eines Problems auf die Qualität der zugrundeliegenden Sensorik zu untersuchen und damit eine quantitative Einschätzung über die notwendigen Randbedingungen im Einsatz dieser Sensorik ermöglicht.

Abstract

In this thesis we are investigating the question of classification of human motion from three different perspectives: Firstly, we develop a methodology that allows the assessment of human muscle signatures to differentiate between gestures they originate from under consideration of natural shifts of those signatures in time with each motion. We can show that our proposed method can leverage Dynamic Time Warping to successfully classify hand gestures based on muscle signals with an accuracy corresponding to or exceeding the current state of the art. In addition to that, we can show that our proposed method allows insight into the underlying mechanisms this classification results are based on and why they are so accurate. We can show that this insight allows for the derivation of intelligent improvements and modifications to the method to improve accuracy even further or to adapt it to specific circumstances. Secondly, we investigate the automated, but understandable derivation of classification procedures for the differentiation between motor patterns inside and outside of intervals of stressful rumination for patients diagnosed with major depressive disorder. We can not only demonstrate that an automated tailoring of a classification procedure is possible, but also that an appropriate splitting between parts of the methodology allows for an assessment of impact for each involved component, yielding the option of intelligent choice for the underlying problem. As a consequence, we can propose a method that is tailored specifically to the underlying structure of the foundational patient data that was obtained within the project and is capable of yielding a lightweight, optimized classification procedure that is also suited for medical applications due to its understandability. Thirdly, we investigate the question of optimal sensor location, to derive an optimal sensor layout for specific or comparable applications under consideration of application-specific constraints and requirements. We propose a method based on an exemplary reconstruction problem that allows the automated assessment of quality of a specific sensor layout. This assessment is then shown to be useful as a basis on which an automated procedure can derive the optimal sensor layout for a specific application. In addition to that, we can show that this method is capable to quantify the sensitivity of the underlying problem to properties of the sensor hardware in use, allowing for a quantitative assessment of the necessity of constraints on the sensor hardware used.

Acknowledgements

I would like to thank Katja Mombaur for the unique opportunity to work with her and on this project, for the creative freedom for the particular questions of research and for valuable feedback, help, insight and support and for providing a great environment to aspire to yet uncharted territories of research.

Furthermore I would like to thank ORB for great colleagueship, in particular Anna-Lena Emonds and Kevin Stein for making the time at our office so enjoyable, as well as Alexander Schubert and Sarah Englert, Catherine Proux and Sabine Volk for invaluable administrative support and Wolfgang Stumpfs for our small, but pleasantly unreasonable projects.

I would like to thank Knut Schnell, Miriam Stein and Julia Hasenkamp for excellent collaboration, feedback and assistance, particularly in the areas of their expertise which was not mine, as well as the colleagues at Lehner GmbH, in particular Rainer Ochs.

I also want to extend thanks to my family who supported me throughout this journey.

Finally I would like to thank the Federal Ministry of Education and Research in the context of the Mitassist project as well as the Heidelberg Graduate School for financial support during the creation of this thesis.

Introduction and Motivation

Autonomous classification procedures are currently a topic of widespread interest, both in research as well as in a wide range of real-world applications from industry processes over to medical image processing to patient monitoring systems. Especially with the surge in Deep Learning techniques in fields from image recognition, speech recognition, text recognition as well as corresponding media generation, autonomous driving, robot operation or fraud detection, these techniques have demonstrated significant capabilities and improvements for these fields compared to previous technology used for this, allowing for example for autonomous text recognition and translation within a smartphone on the go or advanced driving assistance systems that have prevented numerous accidents in the past.

That being said, we also find a tendency of using these learned models with little thought for the underlying mechanics that lead to the obtained classification result. Besides a lack of functional understanding, this comes with the severe consequence of reproducing defects and biases that stem from the underlying data, which is particularly dangerous if the resulting system is used to make decisions on someone's behalf. The fact that we have seen systematic discrimination in terms of image tagging or even loan decisions is just one of the warning signs that a lack of understanding of such systems is dangerous in many ways. After all, if these systems already fail at this point, it is difficult to predict what kind of failures will occur or have already occurred in practice where the consequences cannot be externally checked. This is particularly true in the field of medicine Therefore, some publications explicitly call for understandable classification systems [96]. If used correctly, such informative systems could improve the understanding of pathological mechanisms and promise significant improvements as they can yield significant advances in multiple aspects of, for example, therapy as well as treatment in the form of assistive technologies if leveraged appropriately.

Therefore, this thesis will investigate the question of understandable classification, by developing classification techniques that are assembled such that they can be disassembled and understood, while they can still being fed with data and return classification results, such as it is the case with typical Deep Learning techniques. This way, while each sub-component not only has a specific purpose and can be investigated, vetted and understood independently, this procedure remains either entirely optional depending on the application or can still be done afterwards, when the system has shown to yield promising results. It also allows to intelligently slim down the resulting procedures. While typical Deep Learning models today come with millions of degrees of freedom and according requirements to both computational complexity as well as memory, distilling the relevant components can yield a slimmed version of this procedure that is not only more efficient, but also easier to use and better to understand, and also providing effective pointers to where these methods can be improved most effectively, contributing to more purposeful development.

Lastly, the understandability of these components is not only relevant for research purposes, but also for applications especially in the medical field, as due to the aforementioned bias problems that such systems have exhibited in the past as well as security considerations, a classification system integrated into a medical procedure needs to be explainable by law to be put into practice. Also, having an explainable system aids with patient acceptance, as an nontransparent system may erode trust in the medical procedures that are intended to help a patient and consequentially are potentially entirely rejected.

Consequentially, the techniques developed will be applied specifically to two kinds of problems: The recognition of hand gestures and the utilization of motion classification to identify dysfunctional cognitive states in the context of major depressive disorder.

While gesture recognition has been performed before, we specifically want to address the question of understanding the emergence of classification performance to allow both for a better fine-tuning of the method as well as a more straightforward procedure to assess the suitability of a set of gestures for such purposes in terms of whether these gestures are well differentiable. This shows the clear benefit of understandable methods.

The identification of dysfunctional cognitive states on the other hand poses specific challenges for the application in the medical field. To be able to use such a state classification, explainability is paramount for both regulatory reasons as well as for reasons of patient acceptance, posing a unique challenge to the classification problem. We will address this by developing an approach that allows to understand the underlying mechanics and therefore is usable in the context of such an application.

The majority of evaluations and visualizations in this thesis have, unless denoted otherwise, been performed using NumPy [36], SciPy [94], Matplotlib [50] and Pandas [70, 97].

Outline and organisational structure of this thesis

This work is organised into **three parts**, constituting three directions of research:

In the **first part**, we investigate opportunities of gesture classification leveraging muscle signatures in the framework of autonomous, user-operated signal acquisition with a specific emphasis on the consideration of time-shifts of characteristics within muscle signature signals relative to each other. Also, we investigate the impact of of sensor positioning due to self-operation of such hardware by users. We will derive a classification procedure that utilizes the acquired muscle signature signals and show in detail that this classification procedure is capable of achieving close to no misclassification under consideration of distortions of a sample motion in two samples relative to each other. We will furthermore investigate the distribution of muscle signatures with respect to each other to understand the underlying spatial structures that allow this classification procedure to work reliably and provide an understanding of how this applies to reliable identification of muscle signatures.

This will also yield a criterion to select an optimized set of gestures that shall be differentiated by such a method, so that classification robustness can be tailored to the user's discretion, as well as providing options to increase robustness of this classification procedure at the expense of sensitivity, if a user desires to do so. In this context we will also investigate the transferability and individuality of muscle signatures, finding that they are indeed highly subject specific, hampering a robust transfer of reference data between subjects without any form of adaption.

We will further investigate different condensing techniques that aim both at a reduction in computational cost, as well as a robustification of the resulting classification procedure, by introducing a scheme to both identify and generate optimal representatives for gesture clusters we find to yield comparable, albeit slightly reduced overall accuracy at high reference sample counts, but still successively retain high accuracies in the limit of few reference gestures, whereas the uncondensed approach becomes less sensitive.

The **second part** investigates motion as an indirect outcome of an underlying condition: the capability of motion data serving as a predictor for dysfunctional mental states for patients diagnosed with depressive disorder. For this investigation, a patient study was conducted within the Mitassist project to obtain motion data that is associated with dysfunctional or non-dysfunctional

cognitive states based on self-assessment of patients diagnosed with depressive disorder. Based on this data, we develop an optimization routine that allows to find the most accurate combination of, among other options, data, data representation, data characterization and classification procedure in the form of a multi-level Bayes optimization procedure.

We will then review the results of these optimizations to determine the effectiveness and contribution of different components towards the accurate classification of the aforementioned acquired participant data with a focus on four participants that individually acquired at least eight hours of measurement data exactly labeled with their corresponding cognitive state within the study. We first review the effectiveness of different classification schemes on these four participants, both overall as well as broken down by scheme. We will proceed doing so for characterization space transformations and correlate those with both patients as well as classification schemes. We will briefly review the contribution of frequency bands and data kinds towards such classification to successively eliminate non-contributing components of the optimization.

We will then proceed to investigate the sensitivity, specificity and precision of our results and investigate, how and where they contribute towards high accuracies and where high accuracies predominantly originate elsewhere, with a particular emphasis on classification schemes.

We then proceed to investigate the transferability of motion characteristics between patients by constructing a shared dataset consisting of eight participants who each contributed at least 3 hours of labeled measurement data over the course of the study. We repeat the aforementioned procedure and compare it to the results we found for the participant-individual investigations with respect to overall accuracy, accuracy specific to classification schemes and characterization data space transformations, as well as sensitivity and specificity readings for both data space transformations as well as classification schemes.

We follow this up by an identification of optimal hyperparameters for the most promising classification routine, finding a overall consistent result.

Lastly, the **third part** investigates procedures to improve existing IMU sensor placements with the goal of developing a straightforward procedure to find a task-specific optimal sensor layout to maximize acquired information, such as for classification of dysfunctional states in patients with depressive disorder.

Due to the ongoing investigation in terms of patient assessment, this procedure is instead developed on a reconstruction problem that we introduce first, including the corresponding reference data that provide a known solution to compare against. Then we continue with the development of a framework allowing the optimization of specific sensor layouts that are used to acquire the necessary measurements such a reconstruction is based on, as well as the necessary parametrizations to optimize it. Based on this, we first investigate the impact of noise onto the reconstruction problem. We will then investigate the impact of a reduced sensor set onto the reconstruction problem and follow up with an optimization procedure to find actual optimal positions for a predefined sensor layouts, to then draw first conclusions on the potential of such a procedure for optimized information acquisition.

We then summarize the results obtained in this work and review their potential and requirements for both application as well as further research.

Key contributions of this thesis

Mathematical core contributions included in the **first part** are the development of methods to asses the distribution of time-dependent signature data, where the time axis is a component open for variability, relative to each other. Further, we have established methods to intelligently con-

dense these distributions to lower complexity whilst maintaining good a good representation of all data that is included into this condensate, yielding an improvement in both performance as well as averagability of sequences of different lengths, both subtle and substantial, under constraints of variability in the time axis as well as yielding methodological approaches to condense such signature clusters further by proposing definitions for boundaries in non-linear non-metric spaces, yielding to an improvement in description of data in non-metric spaces in an understandable form. We have further proposed methods to assess the distances between not only samples, but whole clusters of same-type samples in non-metric spaces, laying groundwork for time-adjusted classification approaches. We have then verified that our proposed methods not only work, but yield state-of-the-art performance while at the same time allow for informed assessment of the quality of a given set of gestures with regards to separability. We have further shown based on our proposed methods that these methods are robust against perturbations in signature acquisition, as well as capable of providing quantitative insight into the individualness of subject-specific recordings. This provides another contribution in the form of a gesture detection system that is not only understandable but consequentially also suited for motion detection in contexts where auditability is paramount, for example medical applications such as patient treatment and is also directly based on physical characteristics, which can be leveraged the medical context.

The **second part** yields mathematical contributions in the field of automated optimization procedures, as we have proposed a fully autonomous procedure that yields classification schemes capable of distinguishing between dysfunctional and non-dysfunctional mental states of participants using motion as a predictor. In contrast to Deep Learning techniques, our proposed method provides a component-specific assessment of the resulting procedure, again yielding understanding of the underlying mechanics that contribute to the classification. We have also shown that our method allows to assess the impact of individual components onto the overall result, allowing for informed and quantitatively justifiable choices for specific components. By deconstructing such a classification procedure into its components, the resulting procedure then results in a more performant adaption to new data as the number of degrees of freedom an be reduced compared to a Deep Learning model. We have also shown that such an approach is very capable of identifying aversive cognitive states in patients diagnosed with major depressive disorder, particularly when individually assessed, but also when analyzed in a cross-patient context, showing both the individuality as well as the existence of general underlying structures in motion in the context of major depressive disorder. Further we have contributed a way a classification configuration can be augmented with a specific focus on desired properties with regards to therapy, such as a focus on the detection of aversive or non-aversive states, depending on requirement.

Crucial mathematical contributions of the **third part** are the derivation of an approach to automatically find mathematically optimal sensor layouts that yield improved information acquisition compared to a manual assessment of sensor layout for inertial measurement units, with the option to extend to different kinds of sensors. To do so, we have proposed a generalized description of sensor systems, exemplarily for inertial measurement units commonly used in motion reconstruction, which is leveraged to both emphasize the impact and relevance of different sensors for an exemplary reconstruction problem as well as providing a fully autonomous constraint optimization routine that will yield the optimal layout for a specific given problem, while only depending on a base motion and without the need to specific tailoring of the procedure to specific motions. This also implies direct practical applications in the context of wearable-assisted treatment of major depressive disorder in the form of problem-optimized sensor layouts that can serve to improve both quality and acceptance of resulting wearable-hardware.

Part I.

Classifying gestures with muscle signals

Hand gesture recognition systems are gaining rapid popularity due to a variety of applications, ranging from hands-free computer interfaces in industry applications over human robot interactions up to controlling exoskeletons or smart homes or communicating with humanoid robots. Approaches based on optical sensors [59, 74, 80, 104] are easy to deploy and use and particularly useful in prepared, well-lit environments as they do not need participant preparation. However, being an indirect technique, they can assess motion, but they lack the assessment of direct muscle activity. In contrast to that, surface-mounted electromyography-sensors (sEMG-sensors) allow to measure muscle activity directly and therefore to build much more portable and embedded sensor systems that are capable of assessing said activity independent of location and orientation, and do not suffer from restrictions of the visual field. In addition to that, they might pick up on intended, but not executed motion as well by means of preparatory EMG activity.

As a significant portion of the muscles required for finger movement is located in the forearm, its muscular signal is indicative for the gesture or finger motion performed by a person, therefore an sEMG sensor system on the forearm can be used for gesture classification. We therefore propose a new method based on these muscle signals, using them as a predictor for hand gestures for robust high quality gesture recognition (Fig. 1.3).

Surface-EMG-based methods for gesture recognition have already been proposed by several publications, among them artificial neural network approaches to achieve up to 87% accuracy with a single-channel-sensor when attempting to distinguish 4 different gestures [3]. A composition consisting of a Bayes classifier and a k-Nearest-Neighbours classifier can achieve up to 93% accuracy for 4 different gestures with one channel [56]. An increased number of 5-channels for the sEMG sensors combined with acceleration data can achieve accuracies around 98% using a decision tree constructed of several types of classifiers, but with sEMG-measurements only this yields just 79.7% (average) or 85.8% (median) accuracy [100]. This has been surpassed using an 8-channel-sEMG-Sensor by [55] yielding accuracies above 90%, depending on the number of different gestures to distinguish. 97% can be achieved with gel electrodes with 4 sensors [75], but their practical use is limited due to the electrodes required to be wet. An experiment similar to our method has been proposed by [49] for handwriting recognition, but these experiments were equally performed with wet electrodes that were also specifically placed to target muscle groups, making them difficult to operate autonomously and also targeting a very specific set of motion without estimating the differentiability of a set of gestures.

Furthermore, the classification procedures in these papers do not provide an inherent property to distinguish gestures-to-be-classified from background noise and/or gestures not in the classification pool. Most publications [3, 56, 75] only distinguish between predefined gesture recordings and do not address the issue of distinction of notable gestures from background and arbitrary motion at all. In [100] the distinction of active and non-active periods in the EMG-Signal was proposed, which is capable of distinguishing between gesture and resting position, but does not give a clear distinction between desired gestures and arbitrary motion or undesired gestures. Also, it commonly remains unclear how the set of gestures to be differentiated can be composed to allow for good separability and what contributes to the actual achieved classification, leaving a gap in understanding between classification and fundamental mechanics that is currently still prevalent in many applications of Machine and Deep Learning, hindering straightforward, guided development of such tools.

Our approach uses 8 sEMG-sensors surrounding the forearm. This makes the approach more resilient to misplacement to the electrodes as we acquire a more complete measurement of the muscle activity around the entire forearm. In addition to that, different muscle areas playing together for a gesture can be easily correlated to a characteristic gesture signature. Furthermore, we do not abstract the signal into characteristics, but instead perform the classification directly on the EMG-signals, yielding an intuitive way of measuring similarity as well as defining criteria for it to improve accuracy.

We will first layout the kinds of data we are acquiring as well as their acquisition procedure and discuss the recording hardware in sec. 1.1, 1.2 and 1.3. Afterwards, we will discuss the options of comparing the acquired data, both naively as well as under consideration of time shift in motion characteristics in chapter 2, deriving a method of robustly comparing time series data. We will then discuss classification accuracies achieved by this method as well as a fundamental understanding of why these muscle signatures allow for high accuracy classification and what constraints to place on a dataset to retain this property. We will investigate options to improve performance and evaluate the impact on classification accuracy by the different methods, as well as investigating how they structurally impact the data. We will estimate to what degree muscle signatures can be transferred between different participants as a method to alleviate the necessity to record reference data as well as establishing bounds on the amount of data necessary for achieving high accuracy. Finally, we will derive a modification of the developed procedure to both transfer the preceding findings to a live datastream classification and robustify classification at the expense of detection rate, depending on user preference and use case.

1. Participant-autonomous measurement of muscle signatures

1.1. Acquisition Hardware: The Myo-Armband

We measure muscular activity via surface Electromyography (sEMG). Human muscles emit electric charges upon contracting. These electrical charges can be measured, either inside the muscle via needle electrodes, or on the skin surface via surface-mounted electrodes. An EMG measurement is canonically acquired via three electrodes, two over the span of the muscle that is to be measured and a third apart from it noting a reference potential to correct for electrical base potential. The electrodes then collect the emitted electrical charges that then allow for an estimate of intensity of muscular activity as depicted in Fig. 1.1. As it is unclear which electrode is measuring an arbitrarily emitted electric charge, we find both positive and negative potential measurements. Rectifying these signals will yield the electrode-independent overall signal envelope describing the actual muscular activity signal, which will serve as the basis for our analysis.

Hardware-wise, we recorded data on muscular activity by leveraging the formerly commercially available Myo Armband [67], which is placed on the thickest part of the subjects forearm, as depicted in Fig. 1.3.

The Myo consists of eight surface-mounted electromyography-sensors torodially arranged around the forearm and is worn where muscles are thickest, gathering charges emitted by the underlying muscles upon activity, their three contacts per sensor arranged along the length of the arm, leading to an approximate angle of 45° sensors on the inside of the armband between as depicted in fig. 1.2. This angle is only approximate as it is slightly influenced by the exact arm layout, but the variations introduced by this are negligible for the investigation at hand. The data acquisition frequency of those electrodes internally is 200Hz, the resulting data is transferred to a wireless receiver at a frequency of 50Hz. At the point of actual recording, this yields eight synchronous activity measurements at 50Hz over time, one for each sensor, for the covered muscle parties, resulting in a time series per gesture of dimension $8 \times L$ with L being the number of data points in time of the recorded gesture.

The data acquired by the Myo Armband is transferred to a computer wirelessly, where it is recorded using a Python software that was developed based on a publicly available module [105], but modified to allow for separate recordings of EMG as well as IMU measurements.

It is rectified within the Myo Armband itself prior to transmission and is recorded as such. Furthermore, due to operational constraints, the three electrodes are ordered sequentially along each of the segments of the Myo Armband, as visible in fig 1.2, refraining from placing an explicit potential normalization offsite of the measurement point. The advantage of this construction is a substantially simpler operation for the user, as the arrangement of the electrodes is predefined by the encasing of the device, which will be important for the recording procedure discussed in the next section.

To reduce the impact of further data modification, no further processing was applied to the recorded data, especially no smoothing procedures were performed.

As all following investigations are based on comparative analyses, no renormalization was

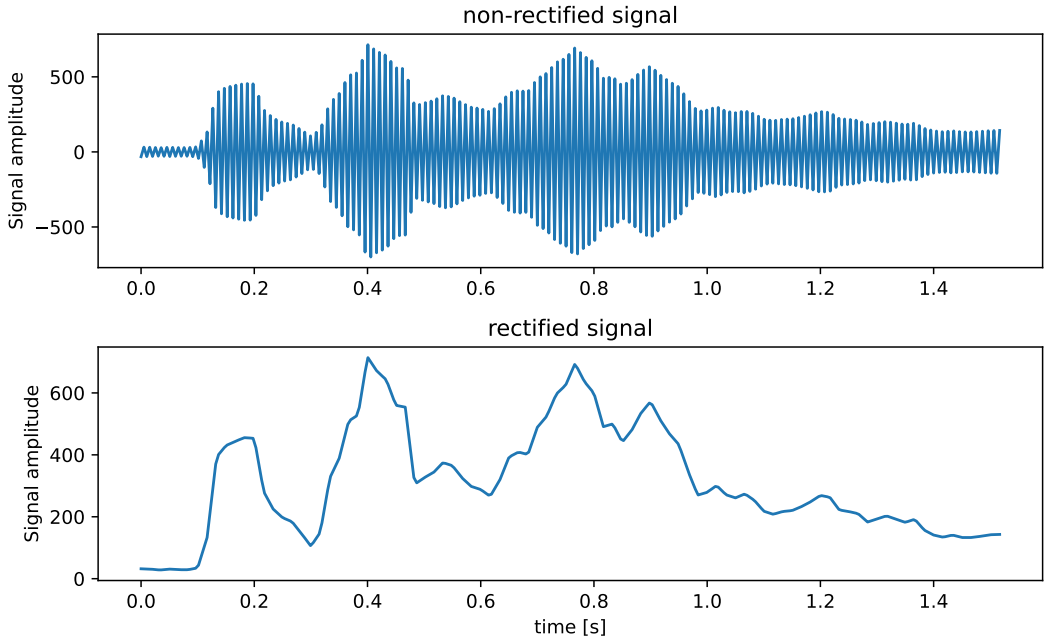

Figure 1.1.: Exemplary electrical signal measured via surface electromyographic measurements. The upper plot depicts a non-rectified signal as it would be acquired by the system whereas the lower plot depicts the rectified signal.

applied to the recorded data to achieve certain general units prior to analysis, as this would not have changed the relation of the samples to each other. That being said, subject-specific renormalization will be investigated later.

Independent of the sEMG measurements, the Myo Armband also provides three-dimensional measurements of linear acceleration, angular velocity and absolute orientation, which were also recorded, but are currently not used for analysis.

Although the Myo armband is discontinued by the manufacturer [58], our approach is independent of this specific hardware, the Myo-Armband happens to provide an easy-to-use sEMG-sensor setup for our measurements.

1.2. Acquisition Procedure

For our analysis we selected seven gestures, also depicted in Fig. 1.3:

1. **fist:** Making a fist from an open hand position.
2. **open:** Opening the hand from a fist position, reverse of *fist*.
3. **count:** Opening the hand from a fist position, one finger at a time.
4. **snap:** Finger snapping with thumb and middle finger.
5. **wave:** Waving in the plane of the hand surface.
6. **wave out:** Waving perpendicular to the hand surface.

Figure 1.2.: The Myo Armband ©Thalmic Labs Inc.

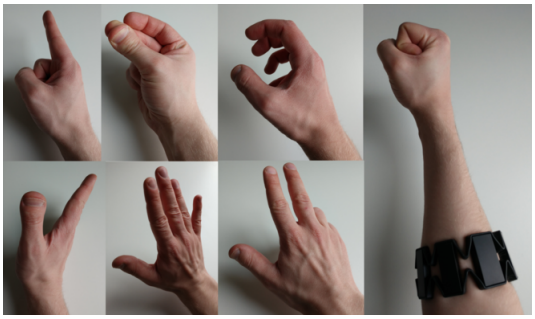

Figure 1.3.: Gestures used to test our classification procedure. Top row, left, to right: point, snap, open. Bottom row, left to right: wave out, wave, count. Right column: fist with Myo Armband attached to the forearm.

7. **point:** Pointing the index finger outwards.

Data has been acquired on two subjects. The subjects were instructed how to seat and orient the Myo armband on their arm and how to perform the set of gestures described above. Furthermore, subjects were supplied with a recording software that randomized the list of gestures to be recorded and showed that to the subject. This list was populated with an equal number for all requested gestures prior to randomizing, ensuring that after recording a predefined number of sequences would be available for every gesture. The subject could then indicate when to start and stop the recording procedure on their own. This allowed participant to record gestures at their preferred pace and timings as well as take pauses of their choice during the recording. Furthermore, the procedure was unsupervised, hence representing a situation of practical applicability, where users of a system based on this methodology would record reference data on their own as well in an unsupervised fashion.

1.3. Acquired Datasets

We used the described setup to acquire several different recordings of electromyographic data on two participating subjects: Firstly, we obtained a bulk recording of 100 measurements per gesture, summing up to 700 samples altogether, without reseating the Myo Armband. This data will in the following be denoted as the “singleday” dataset. In addition to that, we recorded a reduced dataset of 20 samples per gesture and day over the span of multiple days, i.e. with

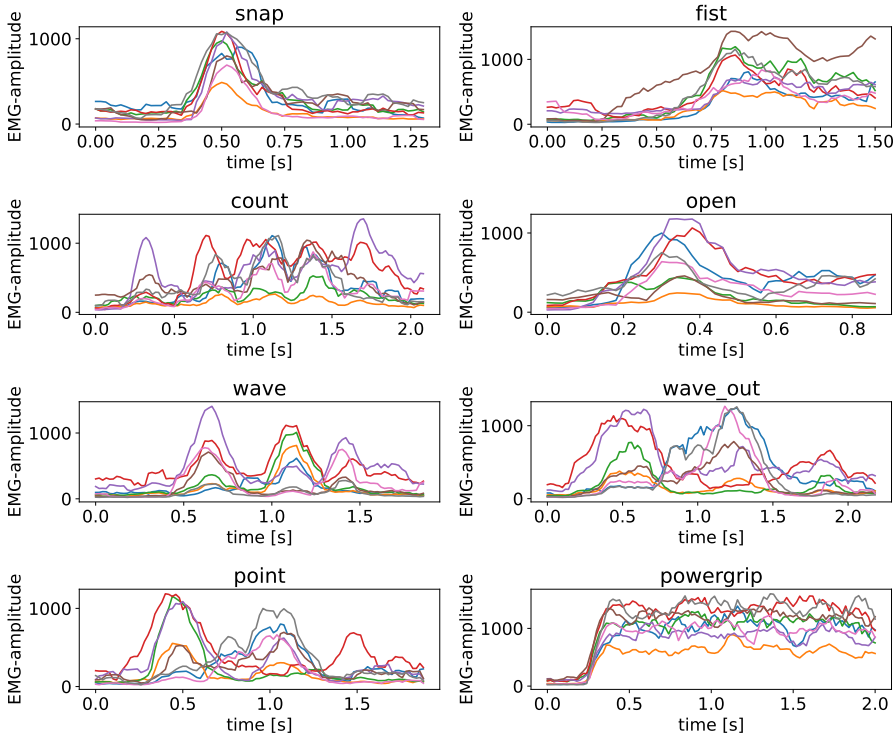

Figure 1.4.: 7 sample recordings, one per gesture with an additional powergrip, each recorded as an 8-channel EMG time series signal. The y-axis' unit is the raw measurement values returned directly by the Myo hardware.

reseating the Myo Armband in the process, consequentially introducing minor random variations in the exact location of the sensors on the wearer's forearm as it would happen in practical usage. A total of 7 days was recorded and the dataset will be generally referenced as the "multiday" dataset.

All recording sessions are identical except for sample numbers, however, to preserve sample number balance (and consequentially an equal weighting of different random exact positionings of the different sensors) the singleday dataset is not merged into the data of the multiday dataset when performing analyses, instead, when compared to the multiday dataset serving as a high-sample-count location-noise-free reference dataset for some parts of this investigation.

Each dataset consists of a 8-dimensional time series per recorded sample, an exemplary excerpt of the acquired data can be seen in Fig.1.4.

2. Assessing the structural relationship of muscle signatures

After acquiring the aforementioned data, we will now proceed to investigate methods of successful recognition of such gestures. To do so, we will investigate the feasibility of a classification system that is capable of recognizing known gestures with high accuracy. Such a system is however, subject to several challenges, in particular when it comes to recognizing non-synthetic inputs, for example of human origin, which are typically subject to a higher degree of variety: Firstly, it cannot be relied upon, that particular characteristics of the recorded time series data are located at the exact same time point for every sample. Instead, it is highly likely that such features are randomly shifted in time over multiple samples. As a consequence, a classification procedure must be capable of resolving such inconsistencies on the time axis. Furthermore, while the datasets described in section 1.3 consist of samples of equal size in the number of sensors used for recording, it does exhibit varying lengths on the time axis, both inherent to different gestures having different common lengths, as well as a certain degree of variety even in samples of the same gesture, as repeated motion is hardly ever identical and consequentially, length of recordings varies, as gestures can be executed at different speeds, but are still considered the same gesture. So furthermore, a procedure in place to recognize such gestures needs to be able to deal with samples of different lengths, both mildly as well as potentially vastly. Lastly, we will investigate how to apply this method to a continuous datastream to identify characteristic structures time-independently.

2.1. Structure-respecting similarity assessment

Be F a feature space. Be furthermore $R \subset F$ a reference population of data points from said feature space F and

$$D : F \times F \to \mathbb{R}_{\geq 0} \tag{2.1}$$

a measure of distance between two points from F. Be $\mathcal{M}$ a mapping identifying all points in R with a class $i \in \mathbb{N}$:

$$\mathcal{L} : R \to \mathbb{N} \tag{2.2}$$

$$r \mapsto i \tag{2.3}$$

We can then employ the k-Nearest-Neighbours algorithm [15] to assess the class of a given datum $d_i \notin R$ by estimating its distance to each datum $d_j \in R$. Its class l_{d_i} is then determined by the immediate surrounding it is spatially embedded into, i.e. the class that represents the majority of its k nearest neighbours for a given distance function. More formally, be J an index set such that

$$j \in J \implies d_j \in R \tag{2.4}$$

Then we can define the set of nearest neighbours of a datum d_i as

$$\begin{aligned} \gamma :=& \{d_j | d_j \in R, j \in J\} \\ \text{s.t. } & |\gamma| = k \\ & \min_J \left(\sum_{d_j \in \gamma} D(d_i, d_j) \right) \end{aligned} \tag{2.5}$$

We can then determine its class $l_i \in \mathbb{N}$ as

$$\begin{aligned} l_{d_i} &= L(d_j) \\ s.t.\ & d_j \in \gamma \\ & \max_j \sum_{d | \mathcal{L}(d) = \mathcal{L}(d_j)} 1 \end{aligned} \tag{2.6}$$

The accuracy of the k-Nearest-Neighbours algorithm does depend highly on the distance measure D chosen. Among the classical choices for vector-valued data samples are

- the euclidean distance: $D(a, b) = \sqrt{a^2 + b^2}$
- the Manhattan distance (of dimension δ): $D(a, b) = \sum_{i=0}^{\delta} |a_i - b_i|$

For trajectories, classical approach of estimating two arbitrary curves' similarity is a sum of point wise differences, such as computing the sum of squared differences, commonly referred to as χ^2 for two sequences of equal length $(x_0, \cdots, x_n)$ and $(y_0, \cdots, y_n)$:

$$\chi^2 = \sum_{i=0}^{n} (x_i - y_i)^2 \ . \tag{2.7}$$

This expression can be decomposed into two components: The point-wise distance function ($f(a, b) = (a-b)^2$) and the aggregate function ($\sum_{i=0}^{n} f_i$), which in combination yields a similarity measure for two sequences of length n.

However, for sequences of different lengths n_1 and n_2 (be $n_1 < n_2$ without loss of generality), χ^2 is undefined, as it either does not consider samples of the sequence of length n_2 in case of $n = n_1$ or does lack samples required for computation in one sequence in case $n = n_2$. Furthermore, assumed $n_1 = n_2$, χ^2 still lacks the necessary framework for considering shifts of relevant features between the two series, which makes it unsuited for appropriately computing similarities between different recorded gestures. Consequentially, we need a both sequence-length- as well as time-adjusted similarity measure.

However, the samples we are comparing are structurally non-identical time series that we assume have variations in their time-axis, which poses two challenges to these kinds of distance metrics:

1. their incapability to compare sequences of different lengths
2. minimal shifts of structures in time will cause significant spikes in resulting distances even though both vectors represent very similar structures

To tackle the aforementioned challenges, we separate the problem into two problems: one of a similarity measure between samples and the second problem as a classification problem given a similarity measure, both of which will be used to address these issues.

2.2. Time-corrected, length-independent similarity assessment of time series

To address these challenges, we will employ Dynamic Time Warping (DTW) for a time corrected, length independent distance assessment. We will first describe the specific flavor of DTW used in this thesis that will accommodate for the time-correction and technical length independence, followed by an assessment of normalization, which is crucial for quantitative, real length independence of the procedure.

2.2.1. Time-shift independent sequence comparison

Originating from the field of speech recognition [66], DTW is a technique to align two time series that are similar in shape by non-linearly deforming the time series in the time axis, hence yielding much better comparability for data with a variable time component.

Be F a feature space and

$$\begin{aligned} x_n &\in F \\ y_m &\in F \end{aligned} \tag{2.8}$$

with

$$\begin{aligned} n &\in [1, N] \subset \mathbb{N} \\ m &\in [1, M] \subset \mathbb{N} \end{aligned} \tag{2.9}$$

Then

$$\begin{aligned} X &= (x_1 \cdots, x_N) \in F^N \\ Y &= (y_1, \cdots, y_M) \in F^M \end{aligned} \tag{2.10}$$

denote two sequences from said feature space and consequentially

$$(x_n, y_m) \in F \times F \tag{2.11}$$

We then define a *warping path* p of length L as a sequence

$$\begin{aligned} p &= (p_1, \cdots, p_L) \\ p_l &= (n_l, m_l) \end{aligned} \tag{2.12}$$

mapping elements from X and Y onto each other, given it satisfies the conditions

$$p_1 = (1,1), p_L = (N, M) \text{ } \textit{(boundary)} \tag{2.13}$$

$$i < j \implies n_i \leq n_j, m_i \leq m_j \text{ } \textit{(monotonicity)} \, . \tag{2.14}$$

The boundary condition ensures that the endpoints of the two sequences are kept as such, whereas the monotonicity condition ensures that sequences do not get warped backwards in time. Furthermore, the intensity of the warping can be constrained by a *stepsize condition* such as

$$p_l - p_{l+1} \in \{(1,0), (0,1), (1,1)\} \, . \tag{2.15}$$

This set of possible stepsize conditions also implicitly ensures monotonicity. This particular stepsize condition also prevents elements of any of the sequences involved from being skipped. It will be used throughout the rest of this thesis.

We further define a cost function

$$c : F \times F \to \mathbb{R}_{\leq 0} \tag{2.16}$$

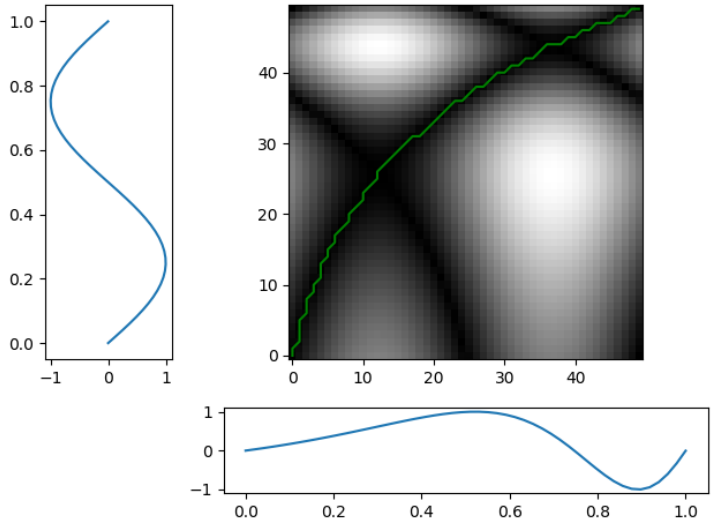

Figure 2.1.: Matrix of point-wise costs of two different time series (cost matrix) with a cost function of $c(x_1, x_2) = \|x_1 - x_2\|_2$. High cost is depicted in white, low cost in black. The optimal warping path, running through the valley of minimal cost, is depicted in green.

that determines a pairwise cost between points of two sequences. We subject it to the requirements of positive definiteness, i.e.

$$d(x, y) \geq 0 \tag{2.17}$$
$$d(x, y) = 0 \Leftrightarrow x = y \tag{2.18}$$

as well as symmetry, i.e.

$$d(x, y) = d(y, x) \ . \tag{2.19}$$

This then allows to associate a cost c_p with each warping path:

$$c_p(p, X, Y) = \sum_{l=0}^{L} c(x_{n_l}, y_{m_l}) \tag{2.20}$$

An *optimal warping path* p^* is then the warping path with the minimal associated cost, i.e.

$$p^* := \min_{(p)}(c_p(p, X, Y)) \tag{2.21}$$

and the DTW-distance of those two series is then

$$c_{p^*} = c_p(p^*, X, Y) = \sum_{l=0}^{L^*} c(x_{n_l^*}, y_{m_l^*}) \ . \tag{2.22}$$

This procedure is shown for two arbitrary sequences and $c(x_1, x_2) = \|x_1 - x_2\|_2$ in Fig. 2.1 including the resulting warping path. A kNN-classifier with DTW as a distance measure has been shown to yield state of the art classification performance on time series data in other applications already [19].

2.2.2. Correcting length-variance

As introduced, Dynamic Time Warping is built to accommodate shifts in the shape of different sequences, realigning motion recordings against one another. As the shape of the constructed cost matrix does only depend on the sequence lengths and the path generation procedure does not impose requirements on the shape of the cost matrix either, Dynamic Time Warping also alleviates the problem of differing lengths, allowing comparisons of sequences independent of their respective number of samples. However, while seemingly alleviated, the problem of differing sequence lengths is actually not resolved, as the resulting similarity measure computed via DTW is still subject to the lengths of the compared sequences.

Theorem 1. *Be $X = (x_0, \cdots, x_n)$ and $Y = (y_0, \cdots, y_m)$ two sequences of lengths n and m. The resulting Dynamic Time Warping distance of a no-skip but otherwise unconstrained DTW procedure of X and Y depends on both n and m within the boundaries of $max(n, m)$ and $n + m$.*

Proof. The DTW distance measure is defined as the sum over the elements of the optimal warping path in section 2.2.

Be X and Y the same arbitrary sequence enumerated by an index set I, i.e. X and Y are perfectly aligned and

$$x_i = y_i \forall i \in I \ . \tag{2.23}$$

This implies, given a defined point-wise cost function $c(a, b)$

$$c(x_i, y_i) = 0 \forall i \in I \tag{2.24}$$

and consequentially

$$p* = ((i, i)_I) \tag{2.25}$$

i.e. the warping path referring only to the elements (x_i, y_i) constituting the diagonal of the cost matrix is optimal. This is equivalent to the sole use of steps of type

$$p_l - p_{l+1} = (1, 1) \tag{2.26}$$

Consequentially, computation of the DTW distance (2.22) of these two sequences reads

$$c_{p^*} = \sum_{i \in I} c(x_i, y_i) \ . \tag{2.27}$$

Hence, the number of summands is $L = n = m = max(n, m)$.

Let now be $X = x_{i i \in I}$, $\#I = n$, $Y = y_{j\, j \in J}$, $\#J = m$, and, without loss of generality, $n < m$, but both sequences still be perfectly aligned despite the different resolution, i.e. sequence Y still describes the same signal as X, just with a higher resolution. Then the warping path p^* will still contain all elements of sequence Y, and elements of sequence X will be matched to multiple entries of Y. This is equivalent to preferring condition (2.26) as much as possible and bridging the remaining elements with stepsizes of type $(0, 1)$. As a consequence, the resulting cost function c_{p^*} will still contain $m = max(n, m)$ summands, establishing the lower boundary of the dependence of the warping path on the compared sequences for different lengths as well.

The upper boundary, i.e. the worst case on the length of the warping path occurs if no diagonal steps are taken in the cost matrix, i.e.

$$p_l - p_{l+1} \neq (1, 1) \ \forall \ l \in [0, L] \subset \mathbb{N} \tag{2.28}$$

$$\Leftrightarrow p_l - p_{l+1} \in \{(0, 1), (1, 0)\} \ \forall \ l \in [0, L] \subset \mathbb{N} \ . \tag{2.29}$$

Therefore, p^* will contain n steps of type $(1,0)$ and m steps of type $(0,1)$ to bridge both sequences X and Y separately, leading to a total number of elements in the p^* of $n+m$, which therefore constitutes the upper boundary.

□

As a consequence, for the normalization factor $\frac{1}{\nu}$ for a DTW-distance between two sequences X and Y of lengths n and m respectively

$$max(n,m) \leq \nu \leq n+m \tag{2.30}$$

should hold. In addition to the two boundaries of ν, we now further introduce

$$\nu = \sqrt{n^2+m^2} \tag{2.31}$$

as an additional candidate for normalization. We can show, that this choice of ν does behave according to our boundaries established in (2.30):

Proof. Be $n, m \in \mathbb{N}$ and, without loss of generality, $n < m$ To show that $\nu = \sqrt{n^2+m^2}$ behaves according to (2.30), two parts need proof:

- $max(m,n) \leq \sqrt{n^2+m^2}$
- $\sqrt{n^2+m^2} \leq n+m$

To do so, we begin with the first part and show (still assuming without loss of generality $n \leq m$) that

$$max(n,m) = m = \sqrt{m^2} \leq \sqrt{m^2+n^2} \tag{2.32}$$

and proceed to the second part, where we show that

$$\sqrt{n^2+m^2} \leq \sqrt{n^2+m^2+2nm} = \sqrt{(n+m)^2} = m+n\ . \tag{2.33}$$

□

This normalization has the specific advantage of balancing the influence of both sequence lengths involved with regard to the resulting DTW-distance. While both $max(n,m)$ as well as $n+n$ only consider extreme shapes of the resulting warping path, namely a warping path consisting of the maximum amount of steps of type $(1,1)$ or $(0,1)$ and $(1,0)$. In practice, however, it is much more likely that the resulting optimal warping path p^* will contain all three possible step sizes, making both normalizations a suboptimal fit. In contrast to that, (2.31) considers both types of steps, single-sequence and double-sequence, thus not only providing much more of a middle ground between the extremes, but at the same time resembling more of a rescaling of the cost matrix to a uniform scaling, making it independent of the involved sequence lengths.

There are further options of normalization one could think of, such as normalizing by path length. This however, would favour more strongly bent warping paths, whereas path-independent normalizations favor less warping and thus penalizing strong warping, i.e. warping that tries to accommodate very dissimilar curves. Hence, for this analysis we stick to non-path-based normalization schemes.

To investigate the use of different normalization schemes, we will in the following compare the three different options, sequence length sum (in the following referred to as *sum* normalization), maximum length of the involved sequences (referred to as the *max* normalization) and the *diagonal* normalization.

As we now have established a distance measure that incorporates the specific length of the involved sequences, which might vary significantly depending on the recorded gesture, we can use this distance metric to define a k-Nearest-Neighbour classification procedure. This procedure reads as follows, for a newly introduced sample X given a set of labeled reference samples Y_i:

1. Compute DTW distance of X to each element of Y_i.

2. Select element Y_i with smallest resulting DTW distance to X.

3. Select label of Y_i as a label for X.

or more formally

Theorem 2. *Be l_x the label of a given sample x. Be $d(x, y)$ a function returning the DTW-distance of two samples x and y. Be further $\{y_i\}_{i \in I}$ a set of reference samples y_i. Then the label of a new sample x is determined as*

$$l_x := l_{y_i} \mid min_{\{y_i\}} d(x, y_i) \forall i \in I \tag{2.34}$$

With this procedure, we can now proceed to classify our given datasets.

2.3. Classifying the *singleday* dataset

First, we perform classification experiments on the cleanest dataset available, which is the *singleday* dataset, which guarantees the same sensor location over all acquired samples. We split the dataset randomly into 70% reference data and 30% test data. We then classify this test data according to the reference data with the aforementioned classification procedure. The percentage of correctly classified samples in the test data (in this particular case 210) divided by the total number of samples (700) yields an accuracy estimate for the method for this particular split. To reduce the impact of the properties of specific splits, this procedure is repeated 100 times to acquire an appropriate statistics over the performance of such a classification procedure. This procedure is again applied to both subjects separately, the resulting accuracies are merged into one set containing 200 accuracy estimates. All of this is repeated for two different normalizations, namely the $1 - norm$, and the $2 - norm$, the results of which can be seen in Fig. 2.2.

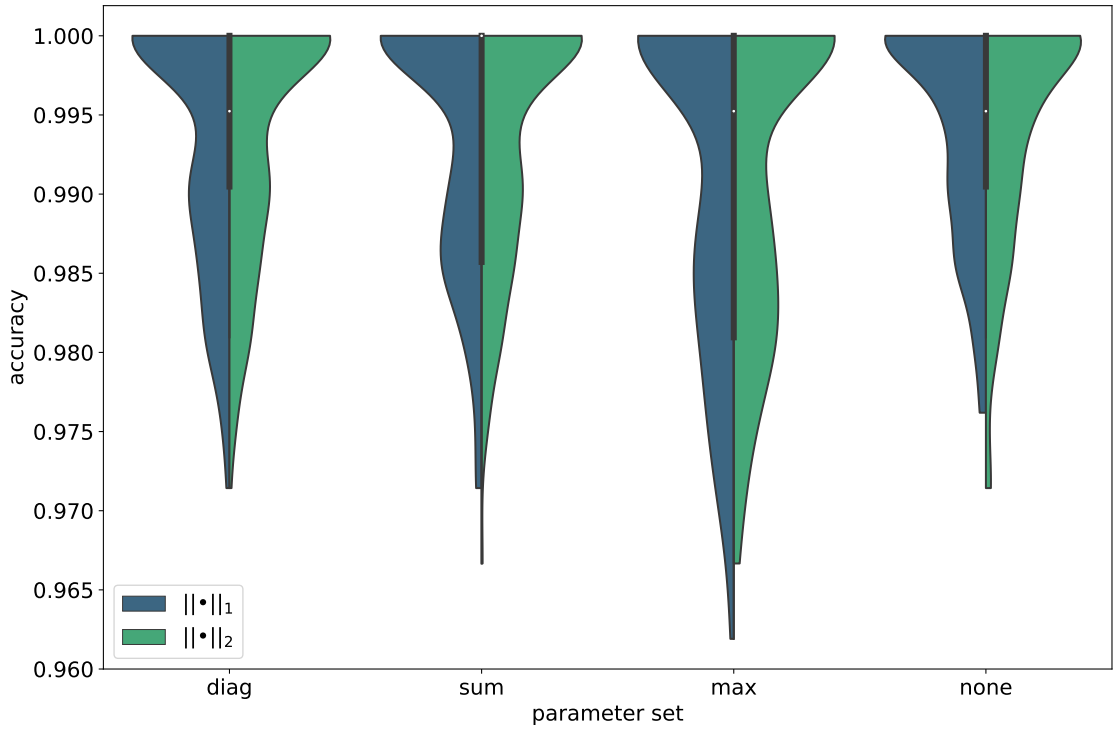

Figure 2.2.: Distribution of accuracy for different hyperparameters of DTW-distance estimate as well as normalization of the resulting DTW distance function of a kNN-DTW-classification procedure. Used are absolute distance and euclidean norm for the DTW-distance estimate and diagonal as well as sum of sequence lengths for normalization. While the surrounding shape indicates the distribution's kernel density, the inner bars indicate the location of the distributions quantiles, namely the white dot the 50th percentile, the thick bar the range between the 25th and 75th percentile, the thin bar extends this range by up to one further interquartile range. The results shown are based on the *singleday* dataset of both participants, each of which was randomly split 400 times into a trainingset with 70% of samples in it and a testset with 30% of samples in it. Depicted accuracies were obtained by classifying the testset data based on a classifier set up with the trainingset data per subject. Resulting accuracies of both subjects were merged for display.

Inspecting the results, we first notice that the accuracies are systematically above an accuracy of 96%, proving the power of the method in recognizing muscle signatures for different gestures. Furthermore, it is apparent that the *max* normalization, compared to both the *diagonal* as well as the *sum* normalization, performs noticeably worse, which may be in part to the fact that it, in contrast to both other normalizations, only considers the properties of one sequence when it comes to normalization, whereas both other methods incorporate properties of both sequences involved. We furthermore see, that the $2 - norm$ produces slightly better results than the $1 - norm$ on this dataset, albeit the differences are rather negligible. This however, underlines the importance of considering both sequences when it comes to normalization, likely due to the fact the the resulting warping path typically describes a curve between the start and end points of the cost matrix as it is also illustrated in Fig. 2.1, rather than being warped in the dimension of one sequence only. Fig. 2.2 also underlines the higher importance of the normalization factor compared to the point-wise cost function, at least with the cost functions tried in this investigation.

2.4. Dataset condensation

While the aforementioned experiment give an insight into the gold standard, the method can provide, it comes with severe disadvantages in the field of computational complexity. In 2018, [30] managed to reduce computational complexity to $\mathcal{O}(n^2/log(log(n)))$, however, this still poses a challenge for a bigger reference dataset, consequentially leading to contradictory incentives: While the enlargement of the reference dataset is generally preferable for higher accuracy, it increases the computational cost, which emerges as an issue for any real-time classification application. Hence, for a real-time application we need to balance these two contradictory goals. As a further reduction in complexity for Dynamic Time Warping seems unlikely, we will now investigate procedures to retain high accuracies when reducing the size of the reference data set.

2.4.1. Selecting representative samples

To reduce the size of datasets of arbitrary size, we aim for finding a representative sample per class that will be used to, as the name implies, represent a class of samples. If applied successfully, this reduces the number of samples per class to one, ideally, without sacrificing too much accuracy.

We define a selected representative $\tilde{x}$ of a given set $\{x_i\}_{i\in I}$ with a given distance measure d as

$$\tilde{x} = x_i \mid R\left(\{d(x_i, x_j) \mid i, j \in I, i \neq j\}\right) < R(\{d(x_j, x_k) \mid i, j \in I, i \neq j\} \tag{2.35}$$

with an appropriate reduction metric

$$R : \mathbb{R}^I \to \mathbb{R} \tag{2.36}$$

which in for the remainder of this series will be the median of a given set of numbers, i.e. we select the sample of a set that has the smallest median distance to all other elements in this set as the representative for this set.

Using this criterion to condensate the reference data of each class down to one representative sample, we can perform a k-Nearest-Neighbours classification, enhanced with DTW for assessing distances, with as little computations as possible with a practical performance increase proportional to the number of reference samples in each class, as this computation needs to be carried out only once and can be performed in advance to any practical application. Classifications performed with this method will be referred to in further analyses by the abbreviation of Rep. However, due to the discreet nature of the problem, the selected representative might still be a suboptimal choice, if the space of available samples is suboptimally distributed, as the representative sample must be within the set of all samples.

2.4.2. Constructing average representatives

To tackle this, we will to compute a representative average sample that better represents the cluster. How to achieve this is not completely obvious, as we cannot simply average the samples, as firstly they might not have the exact same lengths and secondly we risk a similar problem as we have encountered in computing χ^2 on arbitrary series, namely that the actual characteristics of the series are misaligned, which will deteriorate the quality of the resulting signal average, which can clearly be observed in the reduced overall amplitude of such an average visible in Fig. 2.3. Also, while this brief experiment does not introduce additional artifacts such as oscillations, this is a substantial remaining risk.

This therefore requires an averaging procedure tailored towards averaging time series that expects that is capable to account for shifted characteristics of individual series. Such a method, named *DTW Barycenter Averaging* (DBA), is proposed in [28] in the form of barycentering multiple time series leveraging the time-shift correcting properties of DTW.

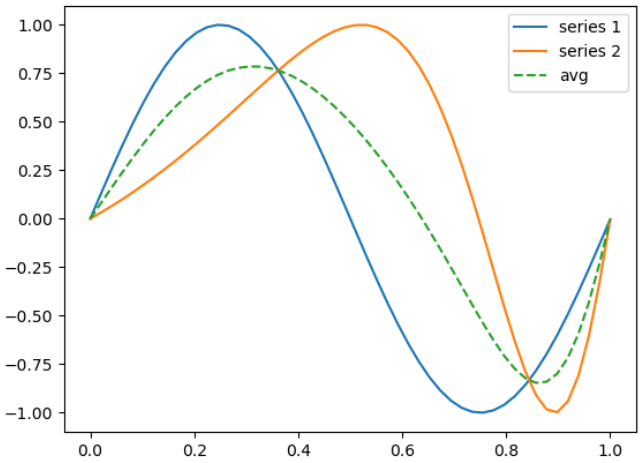

Figure 2.3.: Naive averaging for two similar, but non-linearly time-shifted time series, with a clearly visible reduction in peak amplitude due to the peaks of the series being shifted against each other.

To construct such an improved barycenter averaging procedure, we first need to revisit the derivation leading to (2.20). We notice that this is in fact the sum of the element-wise distance or cost c of two vectors $\tilde{X}, \tilde{Y} \in F^L$, which would, if we would impose linearity onto c, be be writable as a scalar product of $\tilde{X}$ and $\tilde{Y}$ in warped space F^L. To obtain the warped equivalent $\tilde{S} \in F^L$ of a sequence $S \in F^N$, we introduce the *warping operator*

$$\begin{aligned} W : F^{|S|} &\to F^L \\ S &\mapsto \tilde{S} \end{aligned} \tag{2.37}$$

with $L = |p|$ and $p := p(X, Y)$ for two sequences X and Y. Henceforth, if we want to align X and Y, we can write

$$\begin{aligned} \tilde{X} &= W(X, p(X, Y)) \\ \tilde{Y} &= W(Y, p(Y, X)) \end{aligned} \tag{2.38}$$

We furthermore define

$$W(X, Y) := W(X, p^*(X, Y)) \tag{2.39}$$

Given that and writing W as a matrix operator, we conclude that

$$c_p^*(X, Y) = \sum_{i=0}^{L} c\left([W(X, Y)X]_i, [W(Y, X)Y]_i\right) \tag{2.40}$$

Hence, an average of two time series X and Y can be computed in F^L simply via

$$\frac{W(X, Y)X + W(Y, X)Y}{2} \,. \tag{2.41}$$

However, as L varies depending on both time series involved, more than two series do not share a common feature space to be averaged in. Thus, to sum up more than two time series, a shared

feature space $F^{|S|}$ must be chosen and all sequences S_i to be averaged transferred into this space, which can be obtained by picking a representative S from $F^{|S|}$ and g

$$W^{-1}(S, S_i)W(S_i, S)S_i \ . \tag{2.42}$$

This will compute the average of sequence S_i with sequence S in their shared feature space F^L and transform the result back into $F^{|S|}$. Repeating this procedure for different S_i yields a set of sequences in $F^{|S|}$ which can subsequently be averaged. Leveraging this, we compute a barycenter average of a set Θ of similar time series $S_1, \cdots, S_\theta$ by first picking a starting curve $\hat{S} \in F^O$. Then an O-dimensional barycenter average is approached by the k-iteration procedure

$$\hat{S}_{k+1} = \frac{1}{\theta}\sum_{i=1}^{\theta} W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i \tag{2.43}$$

assumed that

$$\exists K | \left\| W^{-1}(S_k, S)W(S, S_k)S - W^{-1}(S_{k+1}, S)W(S, S_{k+1})S \right\| = 0 \forall S \in F^{|S|}, k > K \tag{2.44}$$

i.e. the space unification operators $W^{-1}(S_k, S)W(S, S_k)$ constancify eventually.

Proof. Be $\epsilon > 0$ and $k > K$. We will consider elements of the k-iteration procedure (2.43) in the form of

$$\begin{aligned}
&\left\| \hat{S}_{k+1} - \hat{S}_{k+1} \right\| \\
=&\left\| \frac{1}{\theta}\sum_{i=1}^{\theta} W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i - \frac{1}{\theta}\sum_{i=1}^{\theta} W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i \right\| \\
=&\left\| \frac{1}{\theta}\sum_{i=1}^{\theta} \left(W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i - W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i \right) \right\| \\
\leq&\frac{1}{\theta}\sum_{i=1}^{\theta} \left\| W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i - W(\hat{S}_k, S_i)^{-1}W(S_i, \hat{S}_k)S_i \right\| \\
=&\frac{1}{\theta}\sum_{i=1}^{\theta} 0 < \epsilon
\end{aligned} \tag{2.45}$$

The proposed procedure therefore yields eventual convergence.

□

will use the selected representative as the starting point for a barycenter averaging procedure as described in section 2.4.2 for the set of gestures of each class. The result of this procedure is a generated sample that represents the underlying set of samples much more accurately, allowing for higher classification performance. Classifications with this artificially generated representative will be referred to as *DBA* classification from here on. The original method using unreduced datasets will be referred to as *DTW* for the remainder of this thesis.

Both methods, as well as the unoptimized classification procedure using the entirety of all samples for reference, are tried again on the *singleday* dataset to evaluate their classification performance for different combinations of parameters. The results of this are depicted in Fig. 2.4 The results allow multiple conclusions: The first and most important conclusion that can be drawn from the results is both methods do not exhibit any crucial degradations in classification performance. The results using the selected representative have few trials that show lower accuracies than the unreduced DTW classification as well as exhibiting a lower median

classification result with the sum-normalization, whereas both the classification using the selected representative as well as using the barycenter representations do compactify the accuracy distribution towards higher values. One notable exception to this rule is the max-normalization that actually exhibits worsening accuracies both in the higher and lower extremes as well as the median accuracy when used with the representative selection reduction scheme. While this is to a certain degree also the case for the other two normalizations, the lower extremes in the accuracy distribution are noticeably less common as can be seen in the lower ends of the density estimates for the 1-norm in Fig. 2.4. For the DBA-scheme we find an unconditional improvement in classification accuracy, for the 2-norm even more so than for the 1-norm for this dataset.

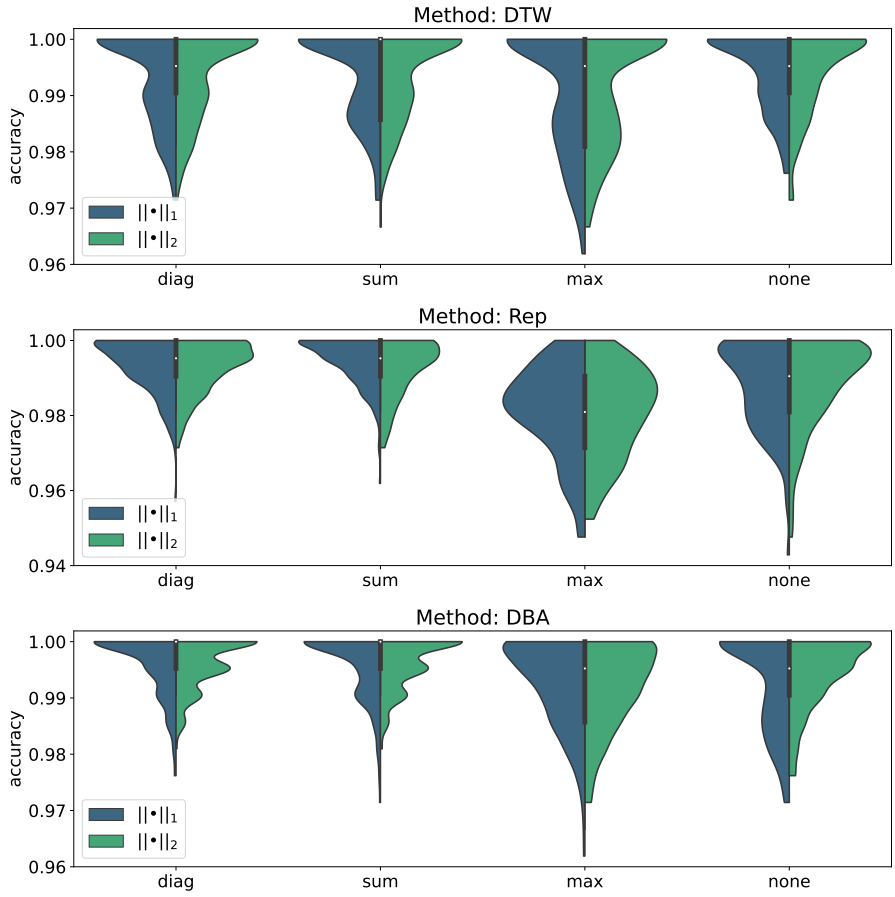

Figure 2.4.: Distribution of accuracy for different hyperparameters of DTW-distance estimate as well as normalization of the resulting DTW distance function of the three different classification procedures introduced in 2.2. Used are absolute distance and euclidean norm for the DTW-distance estimate and diagonal as well as sum of sequence lengths for normalization. While the surrounding shape indicates the distribution's kernel density, the inner bars indicate the location of the distributions quantiles, namely the white dot the 50th percentile, the thick bar the range between the 25th and 75th percentile, the thin bar extends this range by up to one further interquartile range. The results shown are based on the *singleday* dataset of both participants, each of which was randomly split 400 times into a trainingset with 70% of samples in it and a testset with 30% of samples in it. Depicted accuracies were obtained by classifying the testset data based on a classifier set up with the trainingset data per subject. Resulting accuracies of both subjects were merged for display.

2.5. Classifying the *multiday* dataset

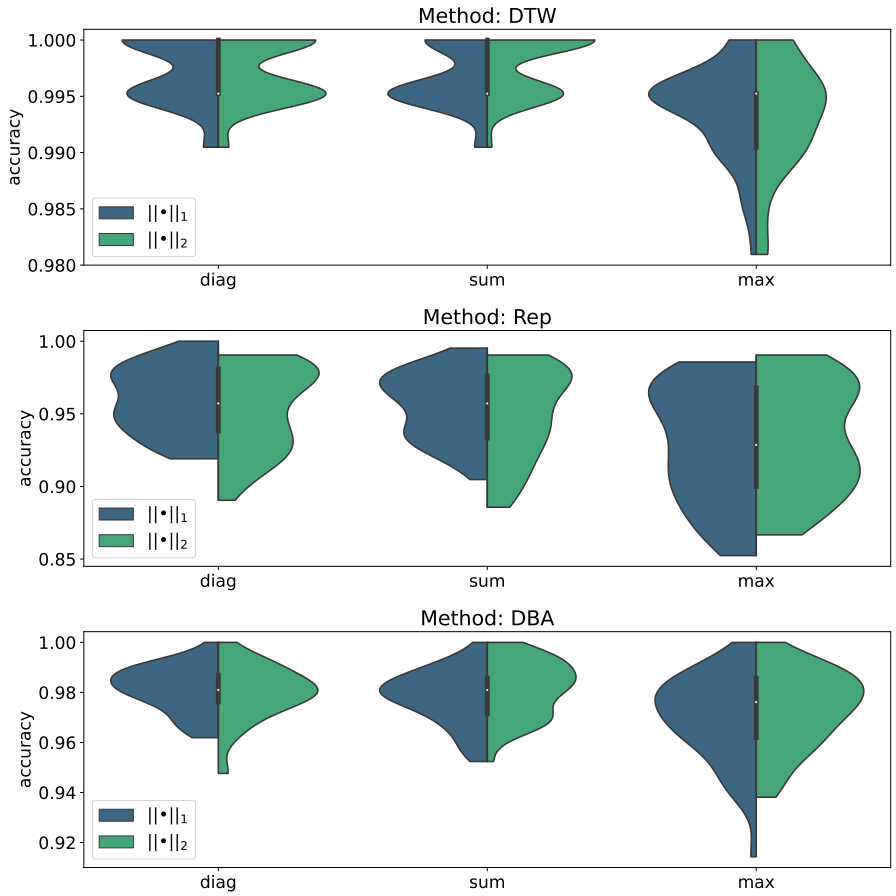

Figure 2.5.: Distribution of accuracy for different hyperparameters of DTW-distance estimate as well as normalization of the resulting DTW distance function of the three different classification procedures introduced in 2.2. Used are absolute distance and euclidean norm for the DTW-distance estimate and diagonal as well as sum of sequence lengths for normalization. While the surrounding shape indicates the distribution's kernel density, the inner bars indicate the location of the distributions quantiles, namely the white dot the 50th percentile, the thick bar the range between the 25th and 75th percentile, the thin bar extends this range by up to one further interquartile range. The results shown are based on the *multiday* dataset of both participants, containing 5 days of data each. The samples of the individual daily subsets of this dataset were merged into one large dataset per subject, each of which was randomly split 21 times into a trainingset with 70% of samples in it and a testset with 30% of samples in it, resulting in a total of 42 data points of accuracy per combination of norm and point-wise cost function. Accuracies were obtained by classifying the testset data based on a classifier set up with the trainingset data per subject, before being merged into one accuracy dataset for display.

To investigate the performance of the aforementioned classification procedure on a dataset with a relaxation on the static sensor location constraint, we repeat the aforementioned experiment on the *multiday* dataset with 500 samples per participant taken from 5 days of recording. The 5 days of data per participant were merged into one dataset before randomly splitting into 70% reference data and 30% test data. Hence, a single such split is not stratified over the separate days, but

this is accommodated by repeated randomized splitting. As this was primarily a verification experiment, the number of randomized splits per participant was reduced do 21, resulting in a total of 42 accuracy data points per combination of norm and point-wise cost function, the distributions of which are depicted in Fig. 2.5. Due to the smaller number of samples constituting the displayed This leads to some quite interesting results: We again see that the diagonal and sum normalizations have a clear edge over the max normalization, however, this time this behavior is not only exhibited in the performance-improved classification approaches, but unmistakably visible in the classical DTW-classification already. Besides that, both the diagonal as well as the sum normalization exhibit a somewhat similar accuracy range when the dataset is condensed to a selected representative compared to the unrefined DTW classification, but the reduction to barycenter average Furthermore, we see that the 1-norm appears to be more resilient to non-static sensor locations for this kind of classification for both the diagonal as well as the sum normalization, while it is actually detrimental to the classification performance when used with the max normalization.

Given the obtained results, we will for the remainder of this analysis focus on classification with the diagonal normalization, as this is the conceptually most robust while yielding comparable performance to the sum normalization, as well as the 2-norm which performed with a slight edge on the *singleday* dataset, to push the developed methodology to its limit.

2.6. The distribution of EMG-signal clusters

In the previous section, we have established that our proposed classification methodology not only works, but also yields very high accuracies. However, so far only qualitative arguments about the "why" have been discussed. In the following, we will quantitatively investigate the conceptual foundations that make this methodology work and allow for a closer understanding of the underlying structures that lead to the aforementioned discussed result. To do so we will first define two measures:

Definition 1. *Be $X = \{x_i\}$ a set of samples of a cluster. Be further $d(x, y)$ a DTW distance measure between two sequences. Then we define the set of intracluster distances of X as*

$$\delta = \{d(x_i, x_j) \ \forall \ x_i, x_j \in X, i \neq j\} \tag{2.46}$$

i.e. the set of DTW distances of each element to each other element in the same set.

Definition 2. *Be $X = \{x_i\}$ and $Y = \{y_j\}$ two sets of samples of two clusters. We then define the set of intercluster distances of X and Y as*

$$\Delta = \{d(x_i, y_j) \ \forall \ x_i \in X, y_i \in Y\} \tag{2.47}$$

i.e. the set of the distances of all elements of X to all elements of Y.

Corollary 1. *For two given sets X and Y of equal cardinality, the cardinality of the intercluster distances Δ between them will be $|X| = |Y|$ larger than the intracluster distances of X and Y.*

Given these definitions, we can associate δ with an estimate for the width of a given gesture cluster, whereas we can leverage Δ to estimate the overall distance between two usually different clusters. To do so, we compute both δ for the samples of each recorded gesture and Δ for each gesture to all other gestures. This allows us to investigate the distribution of clusters with respect to each other in the space that is defined over the given EMG datasets with a DTW-measure for distance, even though this space is not metric. The results of this applied to the *multiday* dataset are displayed in Fig. 2.6 and Fig. 2.7 for subjects A and B respectively. Both figures quantify

what could be concluded from the previous chapter already: We see that for each gesture δ is located around smaller distances than the set $\{\Delta_g\}$ to the other gestures in the set, hence, when leveraging a k-Nearest-Neighbours classification, the results that were discussed in the previous section follow from this finding. However, we also notice that there are substantial differences in the amount of overlap between δ and $\{\Delta_g\}$ between the different gestures: While *fist*, *wave*, *snap* and *count* separate from the other gestures to a degree that the only overlap with any other gesture is beyond the 75th percentile of both distributions involved for subject A (Fig. 2.6), this is not as clear for the three other gestures in the figure. Still, both *open* as well as *point* separate their 50th percentile in δ from the 75th percentile of the other distributions, however, for *wave_out* not even this is the case. Still we can see that still

$$50th(\delta) < 50th(\Delta_g) \; \forall \; g \tag{2.48}$$

holds, which explains the high accuracies found in the previous sections. The distributions of distances depicted also indicate why accuracies in the previous section overall improve when datasets are condensed, as both condensation approaches aim for the smaller half of the respective intracluster distance distributions, hence separating from the remaining samples from other clusters.

For subject B (Fig. 2.7, the results are not as clear, but the same tendencies are visible, especially a proper separation of the intracluster distance median from other clusters. Given the higher overlap here, this is a further indication why classification on the *singleday* dataset compacted towards higher accuracies when condensation was applied.

As a direct consequence, we can now also quantify the quality of the selected gesture set. While it is intuitively plausible that some gestures are better distinguishable than others, this kind of depiction allows to decide which gestures separate well and which need to be reconsidered. In this particular set, we see that *fist*, *wave*, *snap* and *count* are attractive candidates due to their high level of separation from the remaining samples. Furthermore, considering Fig. 2.6, we can conclude that the *wave_out* gesture separates worse than other gestures of this set. Dropping it would elevate the *point* gesture into the choice of well-separating gestures, as that gesture's intracluster distances primarily overlap with the distances to the *wave_out* cluster, already raising the number of well-separating gestures to 5, which is the level that the Myo Armband provides on its own. However, the presented approach allows to easily test further gestures for inclusion into the well-separating set, allowing such a set to be grown to the desired number of detectable gestures successively, including the option to replace already established candidates if they turn out to locate between two otherwise included candidates.

2.6.1. Understanding the average distance disparity between different subjects

A further conclusion from Figs. 2.6 and 2.7 is the overall difference in broadness of both δ and the associated $\{\Delta_g\}$, as well as an overall smaller spread of the different distance distributions. To investigate the reasons behind that, we leverage the additional recordings of a "powergrip" for each subject, which is a recording of the overall maximum voluntary contraction the subject is able to perform. This kind of recording yields the maximum amount of signal that is measurable for a given subject. As can be deduced from Fig. 1.4, powergrips were recorded by starting the recording and then starting the powergrip. Once the powergrip is in progress, we expect roughly constant measurements for all sensors, which approximately holds true for the data shown in Fig. 1.4. The remaining variations can be attributed to both variations in muscle activity unintended by the subject as well as systematic measurement noise due both electronics as well as the skin-mounted, dry sEMG sensors. As these types of measurement noise are both practically

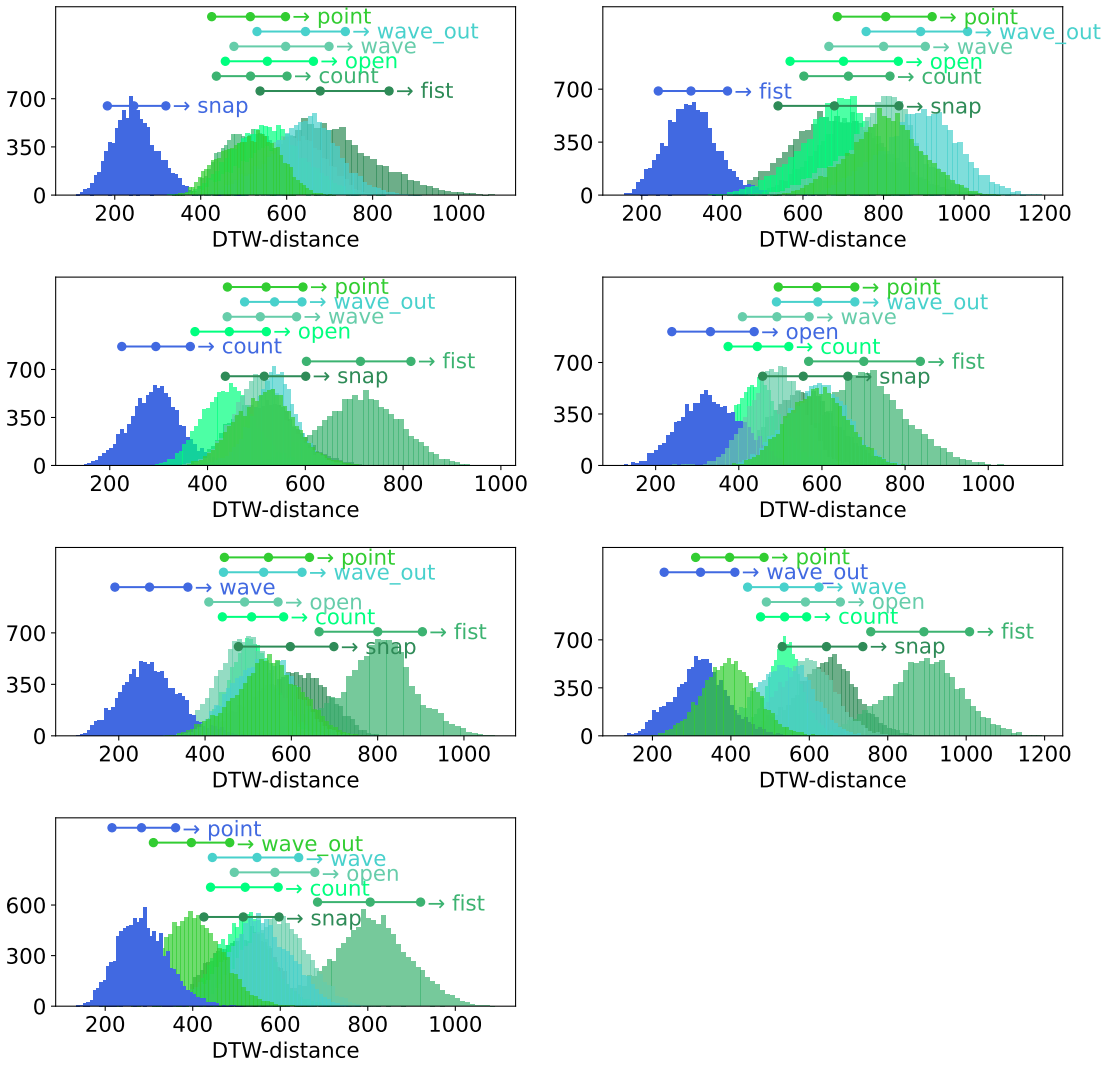

Figure 2.6.: Sample-wise DTW-distances of all samples of a given base gesture to all other samples of this gesture within the *multiday* dataset of subject A, depicted in blue, as well as to all samples of all other gestures present in the dataset, depicted in different flavours of green. Above the displayed distance distribution histograms, each gesture is annotated with a three-point horizontal bar, indicating the location of the 25th, 50th and 75th percentile of the corresponding distribution. The individual days of the *multiday* dataset in question were merged into one dataset before analysis.

separable as well as their separation would not yield significant insight, we will summarize them as measurement noise. To extract powergrip data, data-points are taken from the center third of each measurement. This way the initial starting of the powergrip is removed, furthermore any artifacts towards the end of the recording that could be attributed to immediate fatigue of the subject performing the powergrip.

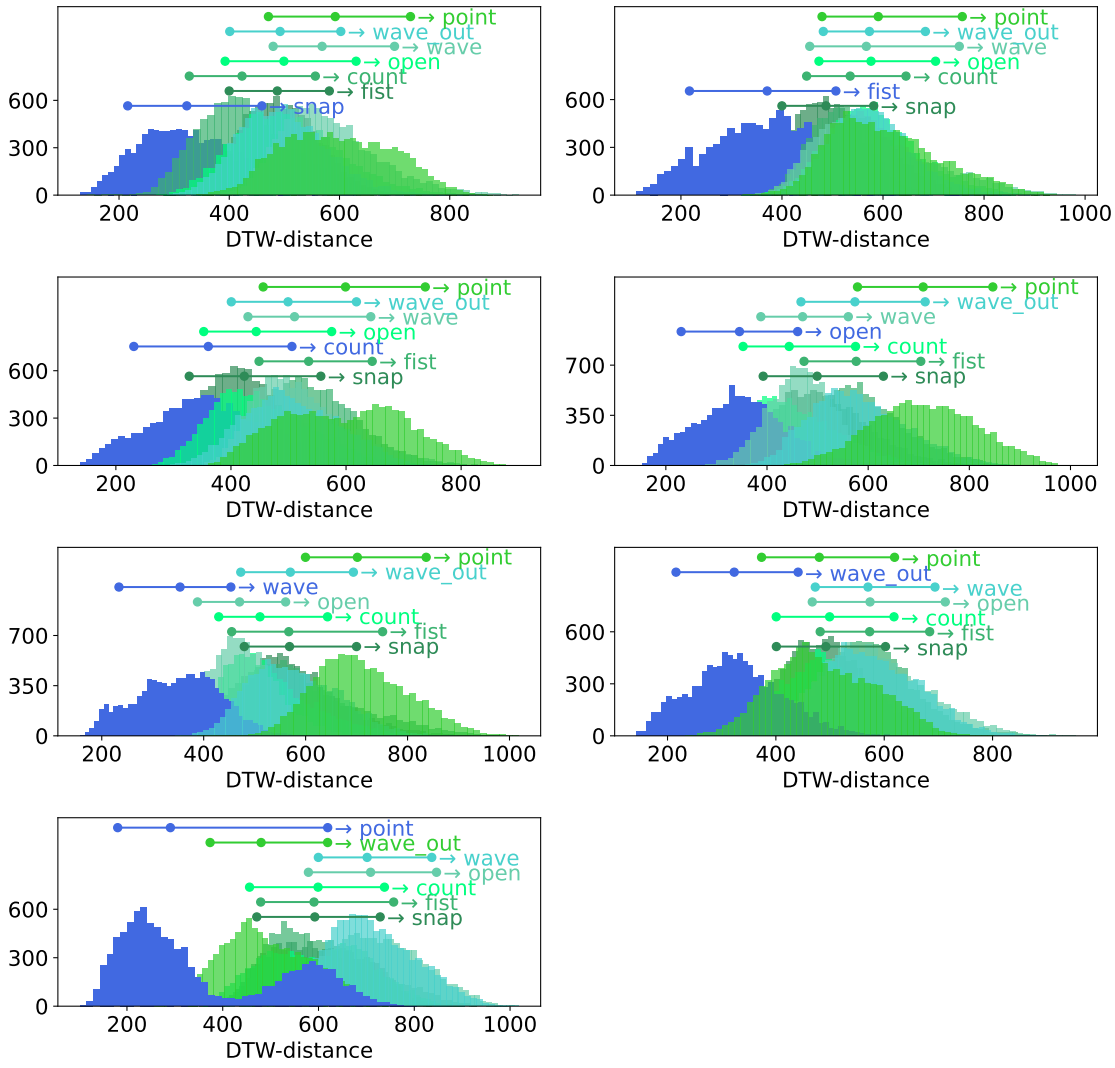

Figure 2.7.: Sample-wise DTW-distances of all samples of a given base gesture to all other samples of this gesture within the *multiday* dataset of subject B, depicted in blue, as well as to all samples of all other gestures present in the dataset, depicted in different flavours of green. Above the displayed distance distribution histograms, each gesture is annotated with a three-point horizontal bar, indicating the location of the 25th, 50th and 75th percentile of the corresponding distribution. The individual days of the *multiday* dataset in question were merged into one dataset before analysis.

Computing the average signal-to-noise ratio, we find that subject A exhibits an overall ratio of approximately 0.15 while subject B exhibits a ratio of approximately 0.23. To test if this could explain the broadening of the distributions seen in Fig. 2.6 and 2.7, we define a test curve

$$G(x) = \frac{1}{2\kappa^2} e^{(x-\mu)^2} \tag{2.49}$$

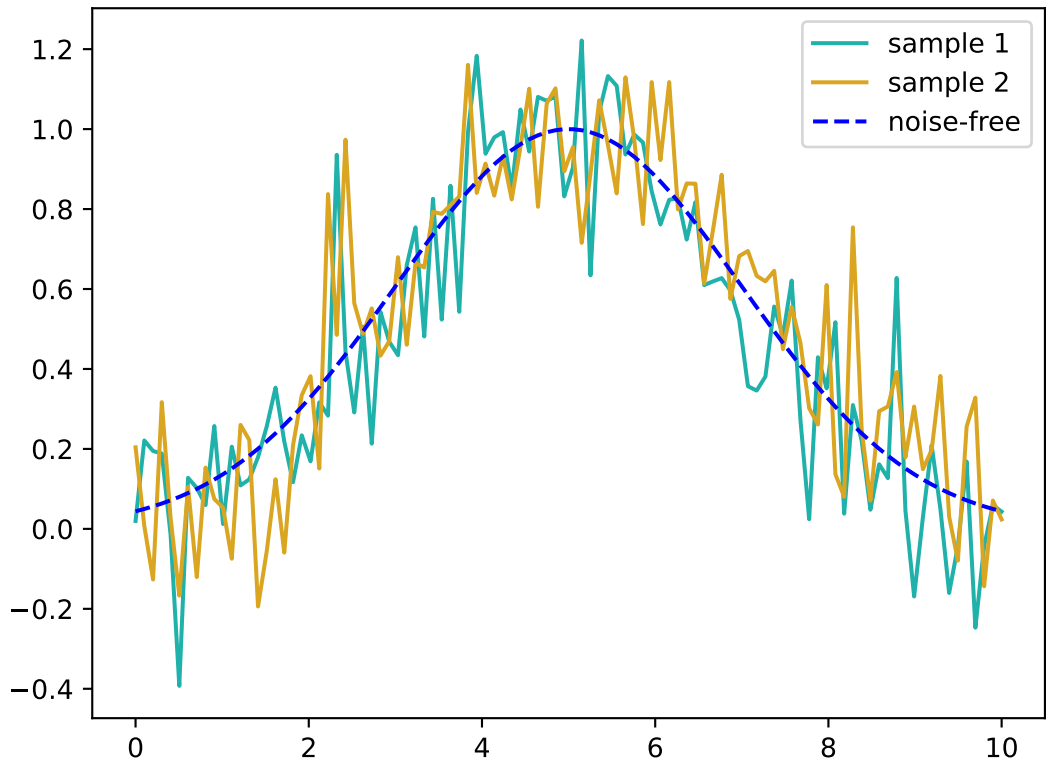

Figure 2.8.: Two samples with different noise-patterns applied to the same noise-free base curve, indicated in dashed blue, with $\mu = 5$ and $\kappa = 2$. The applied noise-pattern was generated by drawing samples from a zero-centered normal distribution with a standard deviation of $\sigma = 0.15$.

and apply random values to each data point. These random values are drawn from a zero-centered normal distribution with a defined width of σ. An example of two of such curves as well as the base curve they are based on are depicted in Fig. 2.8. To estimate the impact of signal noise to the Dynamic Time Warping distance, this distance is then computed on both curves. The procedure is then repeated 100 times for varying levels of σ. The result of this numerical experiment is depicted in Fig. 2.9, in which we clearly see a proportionality of the Dynamic Time Warping distance to the level of noise in the samples used. Unsurprisingly, the variation in Dynamic Time Warping distance increases with the level of noise as well. This, however, does not affect the aforementioned proportionality of the average distance over many different randomizations to the level of noise involved. Hence, as subject B exhibits a higher signal-to-noise ratio by a factor of roughly 1.5, it comes at no surprise that the resulting distances in Fig. 2.7 are both higher as well as broader.

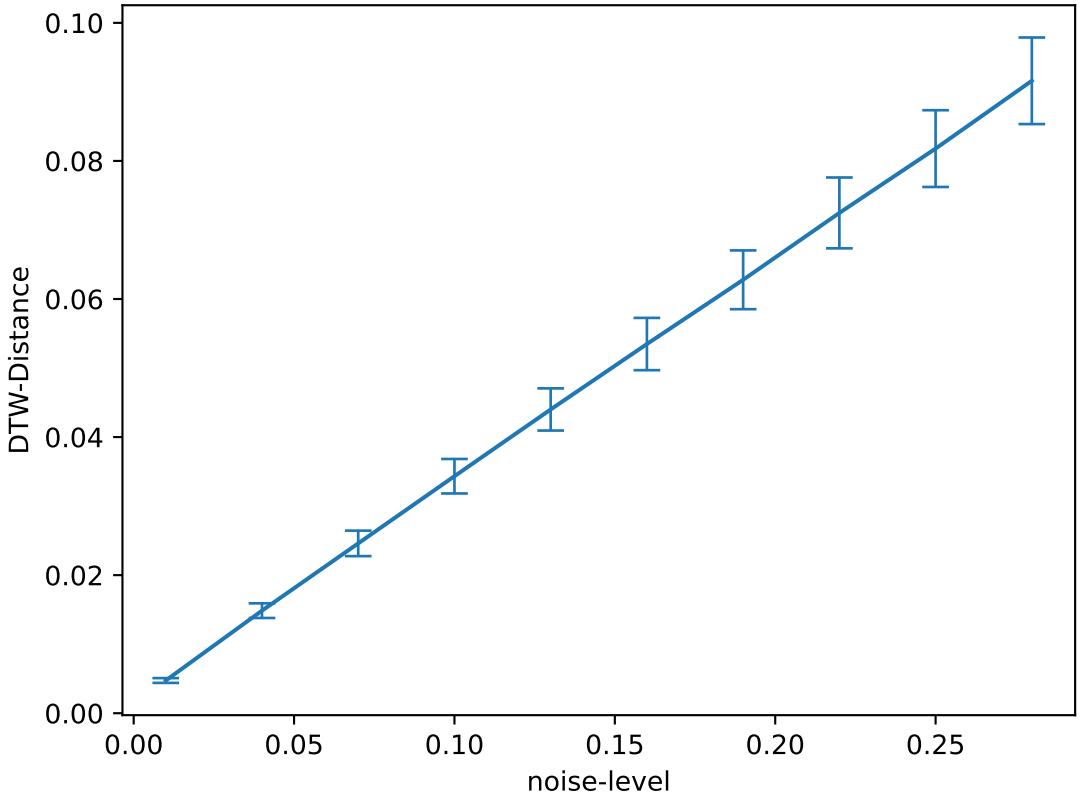

Figure 2.9.: Statistics on DTW-distances between two samples based on the base-curve that was depicted in Fig. 2.8 for different amplitudes σ of the noise. The procedure was repeated 1000 times, the error bars indicate the standard deviation of the resulting set of distances per noise-level.

2.7. Supplementing references via inter-subject data merges

Leveraging the current dataset for constructing a classification system still provides a suboptimal experience for usability, as we currently still require a substantial amount of prerecorded samples. There are two ways of mitigating this problem: Firstly, supplement the given reference data with prerecordings from different persons that ship with such a system, to reduce the requirements on the prerecordings by the user to reach the desired number of reference samples. Secondly, investigate the lower threshold on the requirement of reference samples, so that the number of prerecordings for the user can ideally be reduced accordingly. The second approach will be investigated in the next section, while we will investigate the feasibility external supplementation first, as it is the more attractive one, ideally not requiring any prerecordings of the user for individualization of a gesture recognition system.

To do so, we will compute three properties: We will, for both subjects A and B, compute the intra- as well as intercluster distances per gesture, similar to Figures 2.6 and 2.7, but this time the intercluster distances are between subjects, but of the same gesture, based on the 5-

day *multiday* dataset, which was merged along the subsets into one dataset for each subject. The resulting distances are depicted in Fig. 2.10, which also depicts quite an interesting result: the resulting distribution of distances corresponds to a very clear bimodal distribution for all gestures, with a mostly clear separation of the inter-subject distances. This result is not entirely implausible, given that it is likely that the variances based on the muscular structure between participants contribute more to the variance than variances in sensor placements if they happen to be small enough. Unfortunately, contrasting the resulting distances with Figures 2.6 and 2.7, we observe that the inter-subject distances per gesture are of the same magnitude or higher than the inter-gesture distances within a singular subject. Consequentially, simple supplementation of gestures is unlikely to produce a desirable level of accuracy, especially with dataset condensation methods applied.

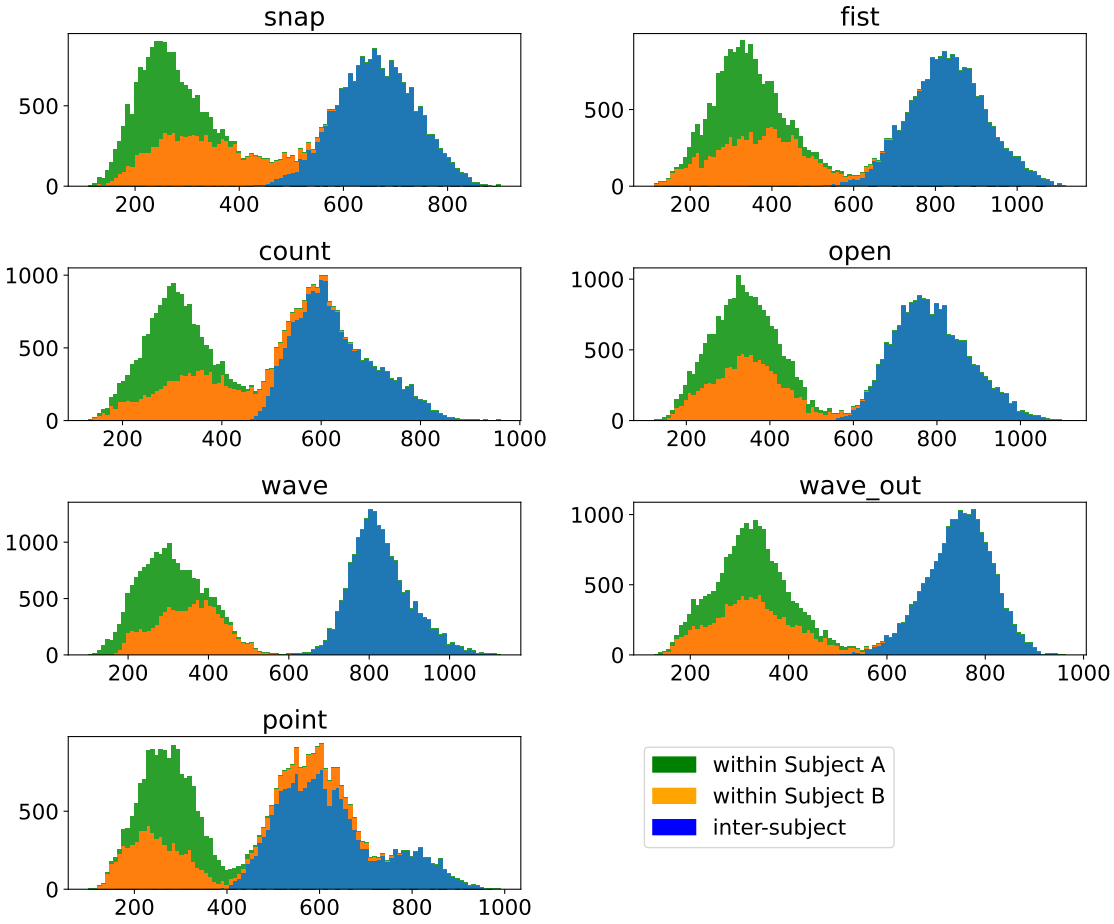

Figure 2.10.: Pair-wise DTW-distances of clusters per gesture within- as well as inter-subject based on unnormalized samples. The within-subject distances are pair-wise distances of samples of the denoted gesture with all other samples of this sample as well as subject. The inter-subject data indicates pair-wise distances between samples of the same gesture, but different subjects. The total histogram consequentially indicates the pair-wise distances per gesture after merging the sample sets of both subjects.

One reason for this can be different physical proportions, especially with regard to muscle strength and consequentially the resulting signal, both in overall intensity, as has been discussed in section 9.1 already.

2.7.1. Data normalization

To adjust for these differences, we will normalize the resulting signals in two ways:

- *batch*: normalize all sensors by an average normalization factor
- *sensor-specific*: normalize all sensors by individual normalization factors

We will, similar to sec. 9.1 extract levels of signal strength per subject from the set of *powergrip* gestures that were recorded alongside the gestures intended for recognition. As the *powergrip* is a way of obtaining values of maximum voluntary contraction, we can use it as a measure of a subject's muscle strength as recorded by the measurement setup in use, and consequentially give us a measure of maximum measurable signal for a subject. From this we can then extract normalization factors for both batch normalization and sensor specific normalization. Again, consistent with 9.1, we use the center third of the available *powergrip* measurements into account, to avoid both starting of the gesture and manifestation of imminent fatigue towards the end of the recording. The resulting averages per sensor are applied for sensor-specific normalization of each sample before re-computation of Fig. 2.10, which yields Fig. 2.12 and the overall average of all data-points of this center third is used as a normalization factor for re-computation of Fig. 2.10 into Fig 2.11. Firstly, we will discuss the batch normalization, whose result is depicted in Fig. 2.11, as it exhibits an number of interesting aspects: First of all, we observe the previously bimodal distribution reshape into a trimodal distribution, where the two intracluster distance distributions separate, while the intercluster distributions retain their own peak as well. The separation of subjects is not surprising: after normalizing both datasets by a constant, the magnitude of the datasets are approximately evened out, however, they still differ in the amount of noise that is present within these dataset. As discussed in section 9.1, subject A already had a lower signal-to-noise ratio compared to subject B. In addition to that, the overall signal amplitudes of subject A are also higher than of subject B. While these two observations are certainly correlated (a higher amplitude at the same level of noise results (by the word) in a higher signal to noise), it also serves to more strongly compress the variance in signal for subject A compared to subject B. So the split we observe can to a significant degree directly be attributed to the proportionality of the Dynamic Time Warping distance to the level of noise present, as depicted in Fig. 2.9.

Whereas such a result is unavoidable, it does not implicate a problem for the merging of multiple subjects into one dataset. Only when observing that the inter-subject distances are still substantially higher than the intracluster distances within the subject with the highest such distances, we clearly see that the introduced normalization procedures do not yield the desired result: enabling us to supplement reference data across subjects. In fact, given that all samples are normalized by the same constant over all samples within a subject, Figures 2.6 and 2.7 depicting the relation of gesture clusters within a subject would only change in their absolute values on the x-axis, but not in shape and relation of the sub-distributions depicted. Hence, it is to be concluded that the distributions of the inter-subject distances are too far away from either subject's intracluster distances to yield any substantial improvement and instead hurting classification performance by diluting gesture clustering.

Repeating this analysis with sensor-specific normalization yields an almost indistinguishable result that is depicted in Fig. 2.12, and consequentially the same conclusions apply. Hence, the approach of supplementing reference data over subjects seems unlikely to be a viable option forward.

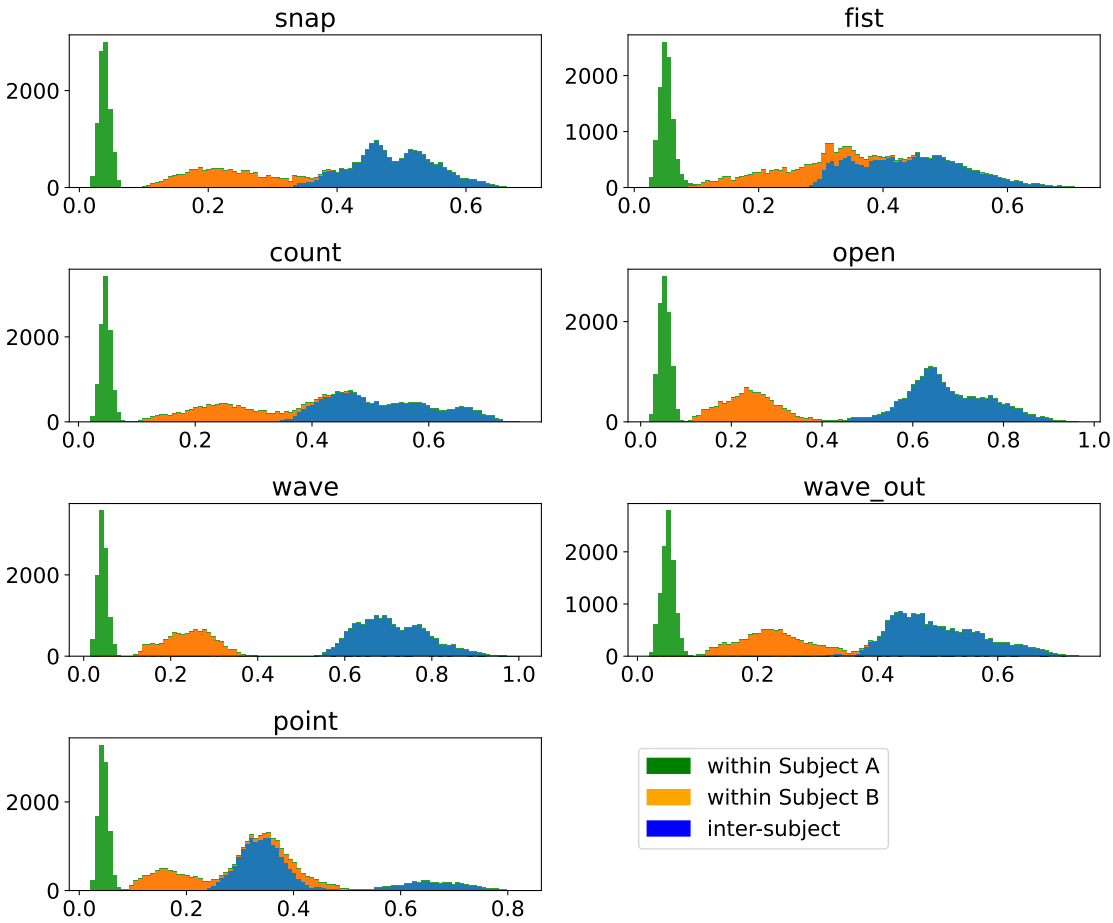

Figure 2.11.: Pair-wise DTW-distances of clusters per gesture within- as well as inter-subject based on batch-normalized samples. The within-subject distances are pair-wise distances of samples of the denoted gesture with all other samples of this sample as well as subject. The inter-subject data indicates pair-wise distances between samples of the same gesture, but different subjects. The total histogram consequentially indicates the pair-wise distances per gesture after merging the sample sets of both subjects.

2.8. Constraining the requirements on reference data

To this point we have analyzed the full dataset of both participants, which was intentionally aimed to be on the high end of the numbers of samples included to aim for a more variability-resistant analysis. However, practical considerations dictate that it is unlikely that a system requiring the recording of 500 gesture samples will gain traction with a user base due to the relatively exhaustive setup procedure. Hence, we will propose a effort-reduced procedure to achieve a good base set of reference data for such a classification. We will investigate the effectiveness of recording a reduced number of samples per gesture over the course of four different sessions. A session is defined by removing and subsequently reequip the Myo Armband, for example over a course of four days, performing one recording session each day. The resulting datum of interest

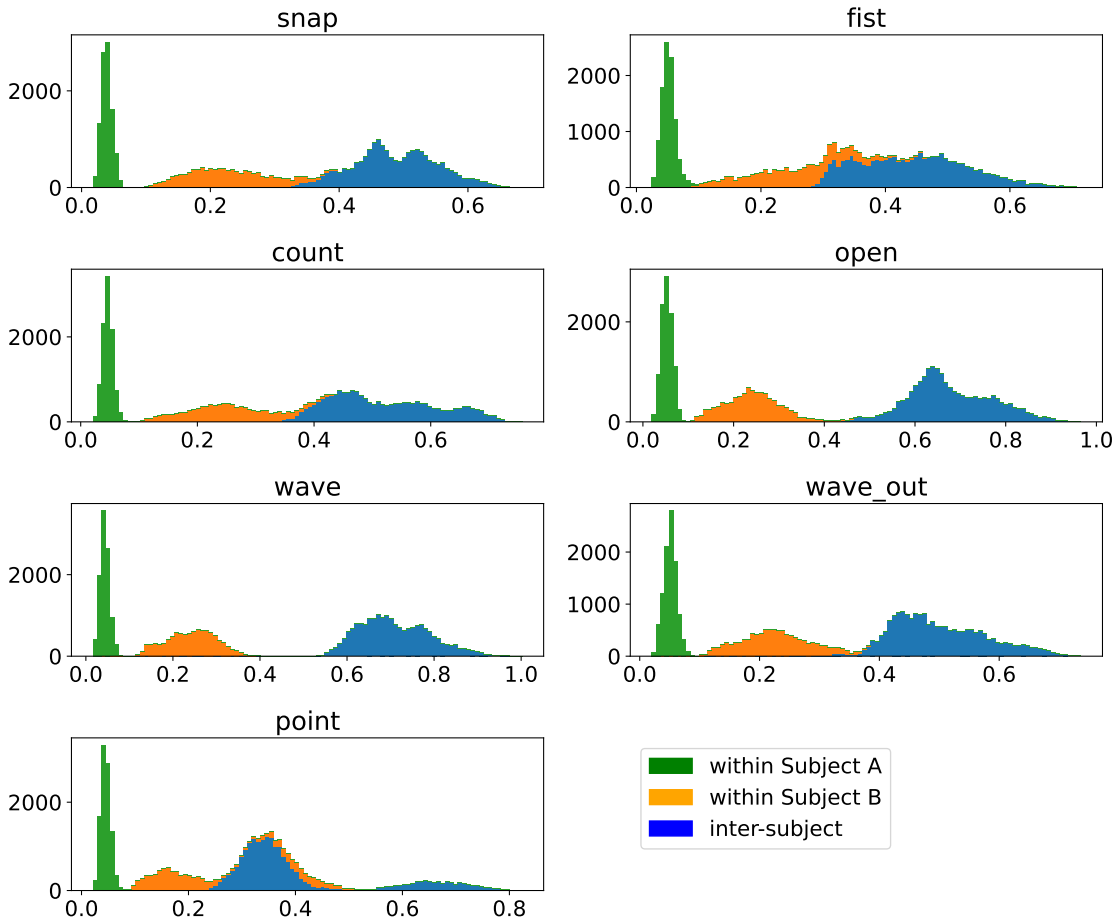

Figure 2.12.: Pair-wise DTW-distances of clusters per gesture within- as well as inter-subject based on sensor-specifically normalized samples. The within-subject distances are pair-wise distances of samples of the denoted gesture with all other samples of this sample as well as subject. The inter-subject data indicates pair-wise distances between samples of the same gesture, but different subjects. The total histogram consequentially indicates the pair-wise distances per gesture after merging the sample sets of both subjects.

is the necessary amount of gestures that need to be recorded per session to achieve an acceptable classification accuracy.

To test this, we will leverage the multiday dataset and will split it into four days that will constitute the set of reference recordings and one day of test recordings. This way, we ensure that the exact seating of the Myo Armband for the test recordings was never exactly identical with any of the reference recordings. This split can be repeated five times, each time another day serves as test recordings, giving us five different configurations of reference and test data. Subsequently, we draw a random subset of $i \in [1, \ldots, 19]$ samples on each reference day, resulting in effective reference sets containing 4 to 76 samples, with a corresponding test set of always 20 samples. This procedure is repeated for every configuration, yielding five different effective reference sets per effective set size, which serve as a basis for a standard classification experiment.

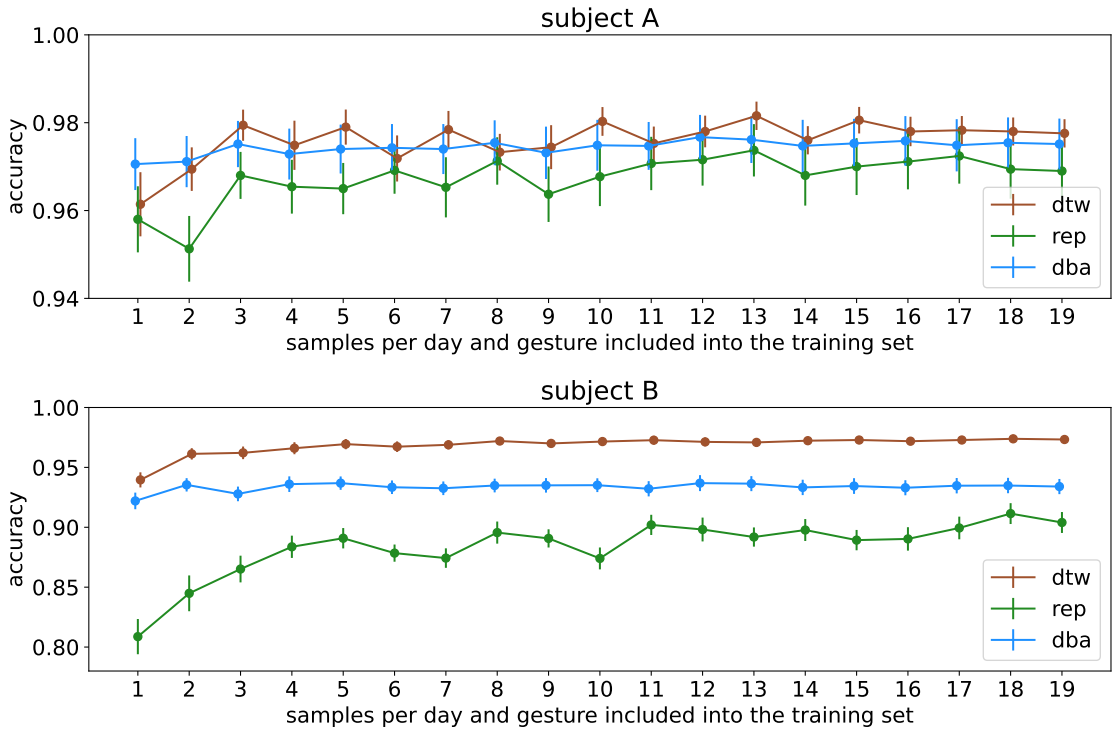

Figure 2.13.: Classification accuracy per subject with different sizes for the training data set. Based on the 5-day-*multiday* dataset, one day was used for the testset, the remaining four days served as the for the corresponding training base set. The x-axis indicates the number of samples that have been split out into the effective training set per day in the training base set, hence the total number of samples in the training set is fourfold this value. The data-points indicate the average accuracy of the different splits, while the errorbar denotes the range of the standard error on this average.

The resulting accuracies, including their standard deviations are depicted for for each effective set size in Fig. 2.13. First of all, the results are consistent with the sub-par performance of the representative-based dataset reduction, exhibiting the worst performance of all methods, with a rather minuscule accuracy gap of about a percent or even less for subject A that exhibited overall high signal amplitudes, but only achieving about 90% accuracy for subject B, which overall exhibited lower signal amplitudes.

We also find that computation on the full set without condensing overall yields the best accuracies, with the exception of the very beginning, where we find an interesting, but also very plausible behavior: At very low sizes of the reference set, we find that the DBA and the uncondensed accuracies are very close, to the point that for subject A DBA accuracy actually manages to exceed the accuracy achieved without condensing, particularly for subject A at a reference set size of 4 or 1 per day this difference is significant. Subject B does not exhibit this particular behavior, but we see that at these set sizes the accuracies of DBA and uncondensed classification are significantly closer to each other compared to larger reference sets. This is very plausible, as the effect of condensing becomes more significant at larger set sizes, as more samples are absorbed into one, whereas at lower set sizes the reduction effect of condensing is significantly reduced due to the smaller uncondensed set, reducing the room for variability between the two.

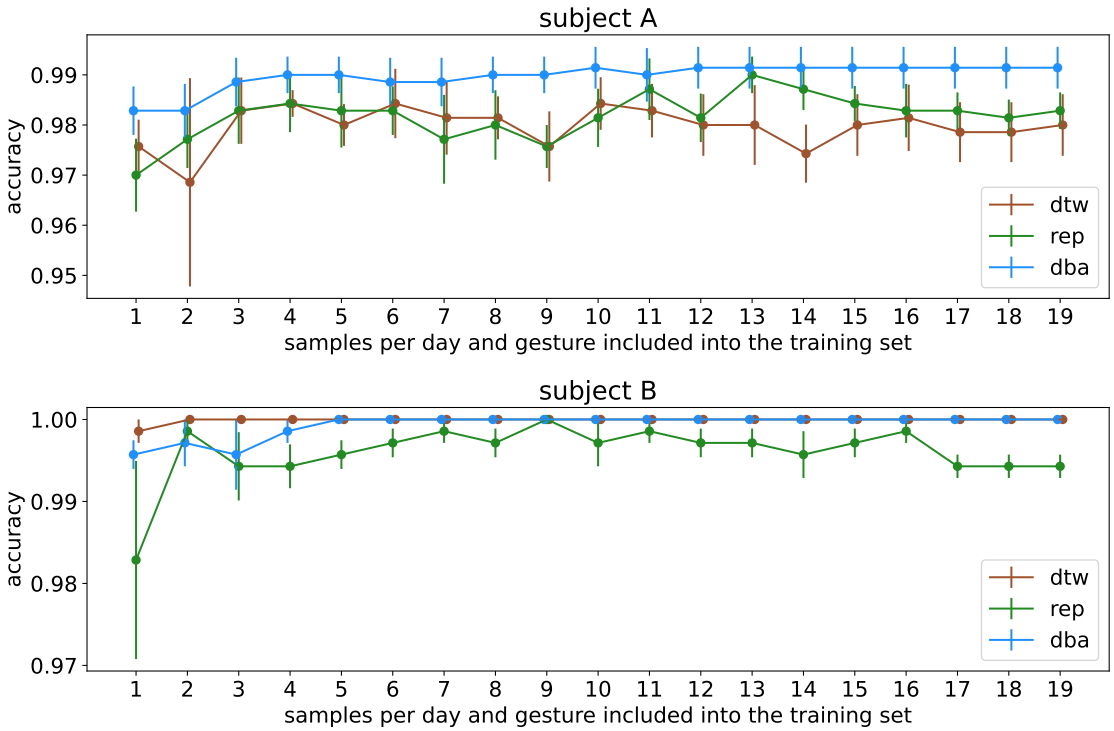

Figure 2.14.: Classification accuracy per subject with different sizes for the training data set for comparison with Fig. 2.13. Based on the *singleday* dataset, the data was first split randomly into 5-day-*multiday*-like dataset, as it is intended as a cross-reference to 2.13, this artificially constructed 5-day-*multiday* dataset is then processed exactly like the data in 2.13.

Yet, we see that there is little variability in total, especially for the DBA-condensed accuracy estimates, which do not exhibit noticeable changes in accuracy with a growing reference set.

This is different for both the uncondensed set and the one computed with a chosen representative, which exhibit noticeably more variations, also degradations in accuracy when growing the reference set, at least for subject A. This effect is less pronounced for the uncondensed assessment for subject B, but very much still present for the representative-selected condensation approach. Subject B also shows a better classification accuracy of about 3 percentage points overall compared to DBA-condensation, whereas representative-selection-based condensation is significantly below that. While the results of subject A are relatively close when considering the variance estimates, subject B shows a clear separation between the different approaches when it comes to accuracy. Furthermore, we do see convergence behavior up to a reference dataset size of up to 5 samples per day and gesture or 20 samples per gesture in total, most prominently observed for subject B and the representative-condensation.

For both DBA-condensation and the uncondensed estimate, convergence behavior is only observed from one to two samples per day or four to eight samples in total per gesture for subject B, for subject A this convergence behavior can be observed up to a size of three samples per day or nine samples in total per gesture, before we observe a saturation in accuracy. Given the slope and variance estimates in accuracy for the DBA-condensation of both subject A's and B's dataset, it could be argued that we do not see a convergence at all in this specific case, but enter

accuracy saturation right away. This tells us two things: First, we see that the DBA is realizing a saturation in accuracies already at very slow reference set already, where the uncondensed set is not sufficient to return a converged accuracy estimate. This is highly plausible, as for very few samples, the fundamental distribution of a gesture cluster might be less than fully represented in the data. A barycenter condensing method such as DBA, is capable of extracting the underlying cluster distribution center better by abstracting away from the cluster, whereas the full dataset might still be too incomplete to properly represent said cluster. This effect is particularly noticeable for well-condensed clusters with high signal-to-noise ratios, such as the dataset of subject A, where we initially see the uncondensed dataset actually performing worse than the barycenter representation of it until an appropriate cluster representation has actually formed. But once this happened, the full cluster data is typically slightly more capable to represent the cluster structure compared to the condensate, but that difference is small and occasionally the resilience of the barycenter condensation to outliers shows as the average accuracy of the uncondensed assessment dips below the one achieved by barycenter condensing, such as with reference set sizes of 6 or 8 samples per gesture per day or 24 or 32 samples per gesture in total respectively. Also, considering the variance estimates on each sample, barycenter condensing and no condensing are barely distinguishable statistically.

For less advantageous signal-to-noise data, such as that of subject B, we still find advantages of the full cluster representation over barycenter condensing with regards to accuracy, reinforcing the obvious finding that the reduction of a cluster will strip away certain information information. This stripped information, however, seems to be continuously less relevant the clearer and consequentially cleaner the overall signals are.

Returning to our practical considerations, we find that recording two samples a day per gesture for four days (or for seatings of the hardware) seems to be already sufficient to produce a consistent classification result when barycentered. This is both in line with performance estimates, where condensed reference data is substantially more advantageous compared to non-condensed reference data as well as the fact that barycentering has indicated to be advantageous especially at low sample sizes, particularly for participants exhibiting strong muscle signals, but also for less advantageous signal-to-noise ratios, barycenter averaging only takes a minor hit in accuracy while guaranteeing early convergence towards accuracy saturation.

To verify these results we will repeat the exact same procedure based on the *singleday dataset* by randomly constructing an artificial multiday dataset from it that matches the dimensions of the native *multiday dataset*, which we will refer to as the artificial multiday dataset. Afterwards, the exact procedure used for the real multiday dataset is performed, the result is depicted in Fig. 2.14. The fundamental difference is that in this case we omit the reseating procedure, i.e. all samples were recorded with the exact same seating. What we find is highly interesting: First of all, we find the resulting accuracies to be unsurprisingly substantially higher as the data is taken from an inherently more consistent dataset. While the real multiday-dataset ranged from 95% to 98% accuracy for subject A and about 80% to 97% accuracy for subject B, we find accuracies starting at about 97% accuracy for this artificial multiday dataset. This is unsurprising due to the identical sensor placement when recording this particular set of data. What *is* surprising however, is for one, subject B is achieving substantially higher accuracies in this dataset, indicating that their sensor variability when reseating the sensor system is higher, whereas it seems to be lower for subject A, who in fact achieves about the same accuracy without condensing. *With* condensing, however, we find that a selected representative can be brought about on the same level of accuracy, for certain configurations even exceeding it, as the uncondensed assessment, indicating that he poor performance of this condensation approach was substantially dominated by the cluster distribution of the multiday dataset which expectedly does not exhibit the exact same placement for the gesture samples of each recording session, but similar enough placements to still guarantee a sufficient clustering of gesture recordings, whereas

this data removes the variance that is caused by different seatings of the sensor system, leading to better performance. Condensing the data into its barycenter improves on this even further, consistently raising the accuracy by another percentage point, which is especially substantial given that we are at 99% accuracy already, leaving little room for improvement. This again shows the regularization potential of barycenter condensing for clusters, smoothing away pronounced cluster corners.

This result also translates to subject B, where we still see a worse performance when condensing to a selected representative instead of on-par performance with the uncondensed dataset, but with a substantially smaller margin compared to the multiday dataset case. We also see a very substantial dip when only considering one sample per day compared to the remaining performance, reinforcing the logical fact that a very small sample size can hurt the representative quality for the data of such a selected representative. The question of condensing to a barycenter is on par with the uncondensed case here, but we hit the accuracy ceiling of 100% correctly classified samples, hence it remains unclear which variant actually performs better in this case. We do, however, see that the barycenter-condensed data needs more samples to converge to this 100% mark, only meeting it at a dataset size of about 5 samples per day and gesture or 20 samples per gesture in total. This could be an indication that the uncondensed state still contains structural information that is lost under barycenter smoothing. However, given the slim accuracy margins we are seeing, it cannot be ruled out at this point that this could also simply be a data artifact.

Overall, we again see a convergence behavior for both subjects, similarly fast for the uncondensed dataset of subject B, whereas the barycenter condensation now converges at the 4 or 5 samples per day and gesture mark, or 16 to 20 samples per gesture in total, while condensing via a selected representative converges only at even higher dataset sizes. For subject A this behavior is more consistent, reaching convergence for all methods at about three to four samples per day and gesture, or twelve to sixteen samples per gesture in total, although the variability for subject A is still consistently stronger to the point of negating the convergence effect altogether, for example at a dataset size of 14 per day and gesture or 56 samples per gesture in total.

Consequentially, we can conclude that we in fact only need about three samples per session and gesture, assumed a four session recording procedure that can either be spanned over multiple days or handled within one day, as long as the sensor system is properly reseated for every recording session, which substantially alleviates the effort to get this system into a functional state for any individual user substantially.

2.9. Improving classification accuracy by introducing acceptance limitations

So far, we have always assumed that the system would always return a classification for a given sample. However, this might not necessarily be desirable, as it could happen that a motion is ambiguous or comparatively arbitrary. In this case we need a procedure to reject samples should they not really fit any classification. Furthermore, if we want to apply this procedure to streamed data, which is the ultimate goal of the proposed method, a way of rejecting samples that do not belong into any of the predefined gesture classes is required. Therefore we will now define how such a system can reject classification if insufficient similarity is detected.

Similarity for this setup is defined by the DTW-distance between two samples. Hence, we can define an upper bound on the required similarity between a given sample and an associated dataset representing a gesture cluster. Specifically, we will place this upper bound on the cluster sample that would be used for the classification, so either the one condensed sample or for the

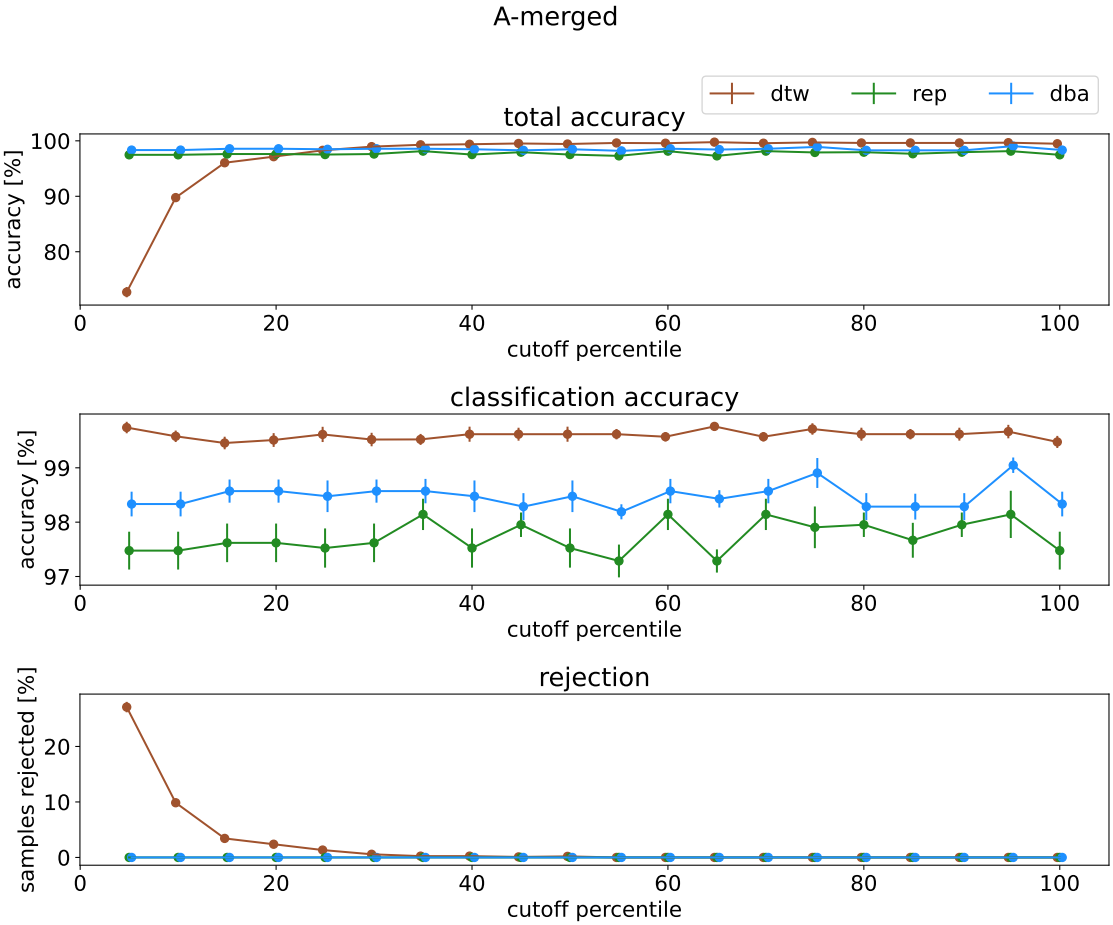

Figure 2.15.: Classification accuracy for different percentiles chosen as cutoff-values for the three classification methods introduced in this thesis for subject A. Classification was performed on the *multiday* dataset, whose individual daily subsets were merged into one large dataset, before being randomly split into a training set containing 70% of the samples and a test set containing the remaining 30% of the samples. This procedure was repeated 10 times. The data-points indicate the average accuracy achieved, the corresponding errorbars denote the standard error on this average value.

uncondensed state, the cluster sample that is closest to the newly introduced sample.

We will again assume the frame of cluster diameters, computing the intra-distances of each cluster, whose largest values estimate the approximate maximum diameter of each cluster. The resulting distribution of intracluster distances will form the basis for an informed estimate for the upper limit of the distance that we still consider acceptable for a class assignment to this cluster if a classification candidate sample would, in the unlimited case, be attributed to this cluster.

To assess the effectiveness of such a limit, we will perform a classification experiment on the more challenging multiday dataset in two flavors: firstly, we will merge the different days of data acquisition per subject into one dataset and then afterwards split this merged dataset randomly into a reference set containing 70% of samples and 30% test set. The split between

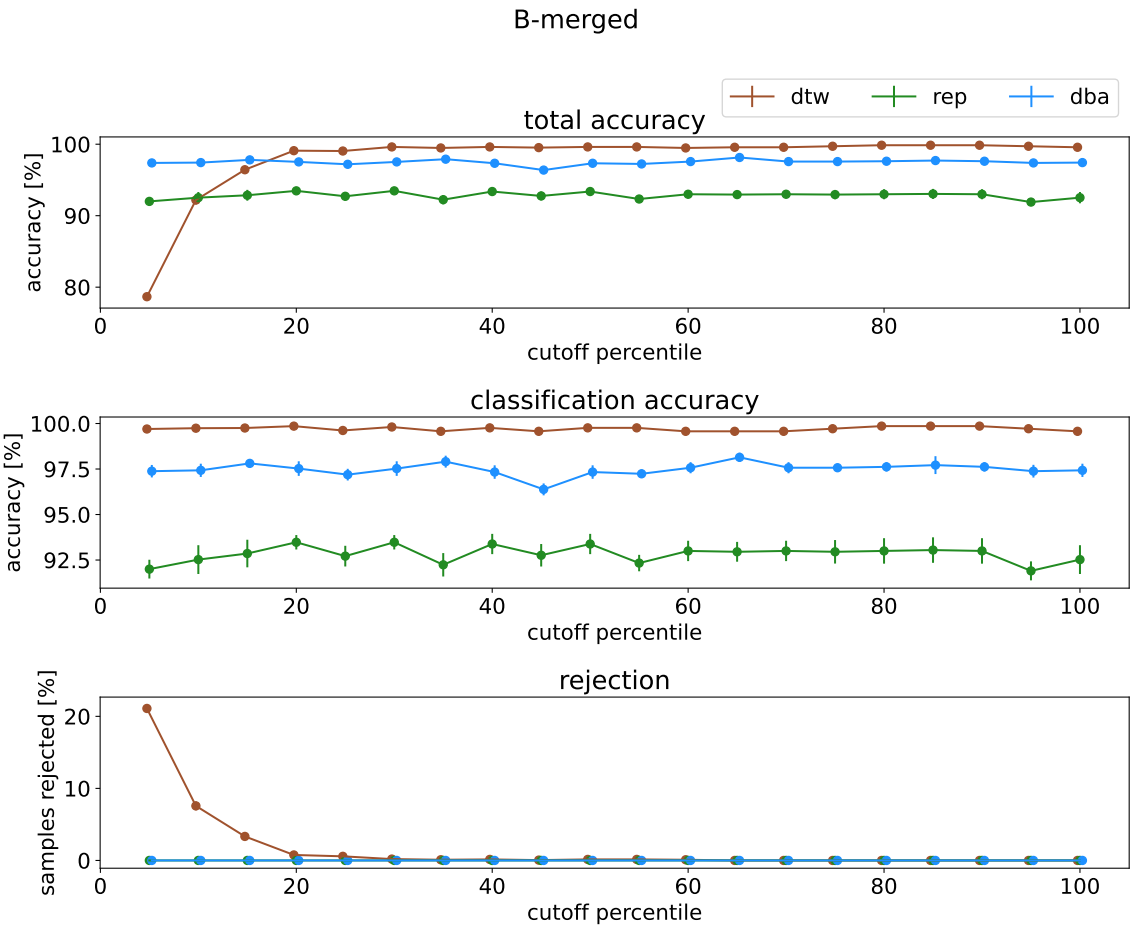

Figure 2.16.: Classification accuracy for different percentiles chosen as cutoff-values for the three classification methods introduced in this thesis for subject B. Classification was performed on the *multiday* dataset, whose individual daily subsets were merged into one large dataset, before being randomly split into a training set containing 70% of the samples and a test set containing the remaining 30% of the samples. This procedure was repeated 10 times. The data-points indicate the average accuracy achieved, the corresponding errorbars denote the standard error on this average value.

subjects A and B will be retained. We will then classify the samples in the test set based on the samples in the reference dataset with different acceptance limits ξ which correspond to different percentiles of the intracluster distance distribution, starting with 5% and increasing in 5%-steps up to 100%, which corresponds to the maximum diameter estimate of the cluster. This means that for an uncondensed classification procedure a classification candidate can have a distance up to one cluster diameter estimate from the outer rim of the cluster constituted by the reference set to still be assigned to this cluster, whereas for condensing techniques a classification candidate could be up to one cluster diameter estimate from the representing sample which approximately corresponds to the cluster center, constituting a tighter effective bound on the acceptance limit ξ. We will discuss two types of accuracies in this scenario: Firstly, the *total accuracy* α_t, corresponding to the fraction of correctly classified samples compared to all samples,

and the classification accuracy α_c, corresponding to the fraction of correctly classified samples among all samples that were not rejected, i.e. had a distance to the cluster they would be assigned to that was smaller than ξ. For each ξ, we will split the data randomly ten times and compute the according properties on each split. Fig. 2.15 shows these properties as well as the percentage of total samples rejected in the process of classification for subject A and Fig. 2.16 shows the corresponding results for subject B. In addition to the absolute values, these also depict the standard error computed over the ten different splits per data-point. We see two fundamental results: Firstly, we find that while the uncondensed case for small cutoffs indeed rejects a number of samples for very tight cutoff distances ξ, this is not the case for the condensing methods. This is highly interesting because it contradicts intuition, as by construction of the cutoff procedure, the uncondensed classification should be more lenient, as the condensing methods effectively tighten the cutoff boundary due to its computation from the center whereas the uncondensed method applies this bound from the outer rim of the reference dataset. However, we see the exact opposite behavior, which strongly emphasises the regularization effect of not only the barycenter average computation, which plausibly represents a more generalized sample of the resulting data, but interestingly already also the representative selection. However, at least at very small rejection bounds, this more aggressive rejection behavior, while being unsurprisingly detrimental to the total accuracy, leads to a higher classification accuracy of the uncondensed procedure compared to the condensed ones. With a relaxing of the rejection boundary, the total accuracy converges towards the condensed methods, eventually exceeding them, while it retains retains an advantage of about one percentage point for subject A and 2.5 percentage points for subject B over barycentering, the generally better of the two condensing methods. This finding, while different in effect for subjects A and B, is overall consistent over both, the differences between subject A and B in behavior correspond to the differences we have already seen previously and can be retraced to according origins.

Secondly, we see that, while the uncondensed method converges with increasing ξ, it does so very quickly, not showing any differences in total accuracy behavior from $\xi = 30\%$ already. This is a valuable insight, because the condensing methods apply ξ from an approximate reference cluster center, whereas the uncondensed approach applies ξ from the outer rim of the gesture clusters in the reference dataset. Therefore, the condensing shortens the effective acceptance limit by approximately half a cluster diameter, hence $\xi = 100\%$ for the uncondensed approach correspond to $\xi = 50\%$ with condensing. However, we do observe convergence in the uncondensed case already at 30%, indicating that the stability the condensation methods exhibit at low ξ are indeed stable, and not just accidential results or dataset artifacts, further reinforcing the benefit of condensing in terms of classification stability as well, even though the classification accuracy is slightly reduced.

In addition to this, we also find that the error on the accuracy estimates is extremely low, particularly for both the total accuracy estimate and for the uncondensed setup in particular, whereas the condensing techniques exhibit a little more variability in case of the classification accuracy, reminding us that we are in fact removing information via condensation, which does have an effect on accuracy stability. Yet, that effect is still small enough to not pose a substantial practical hurdle.

We will now repeat this analysis with a slightly different setup: We will again use the multiday dataset, but this time we will not merge the data into one dataset, but keep the day separation intact and create a reference dataset out of three available days with a test set of the two remaining days. Again, we will recombine the five datasubsets to obtain 10 different configurations for this split, on which we repeat our analysis from before, the results of which are depicted in Fig. 2.17 and 2.17. Overall, we find fundamentally the same results as in the case of the previously merged multiday dataset, with two noticeable differences: Firstly, we find slower convergence for the uncondensed setup, having fully converged at $\xi = 50\%$ compared to $\xi = 30\%$ previously,

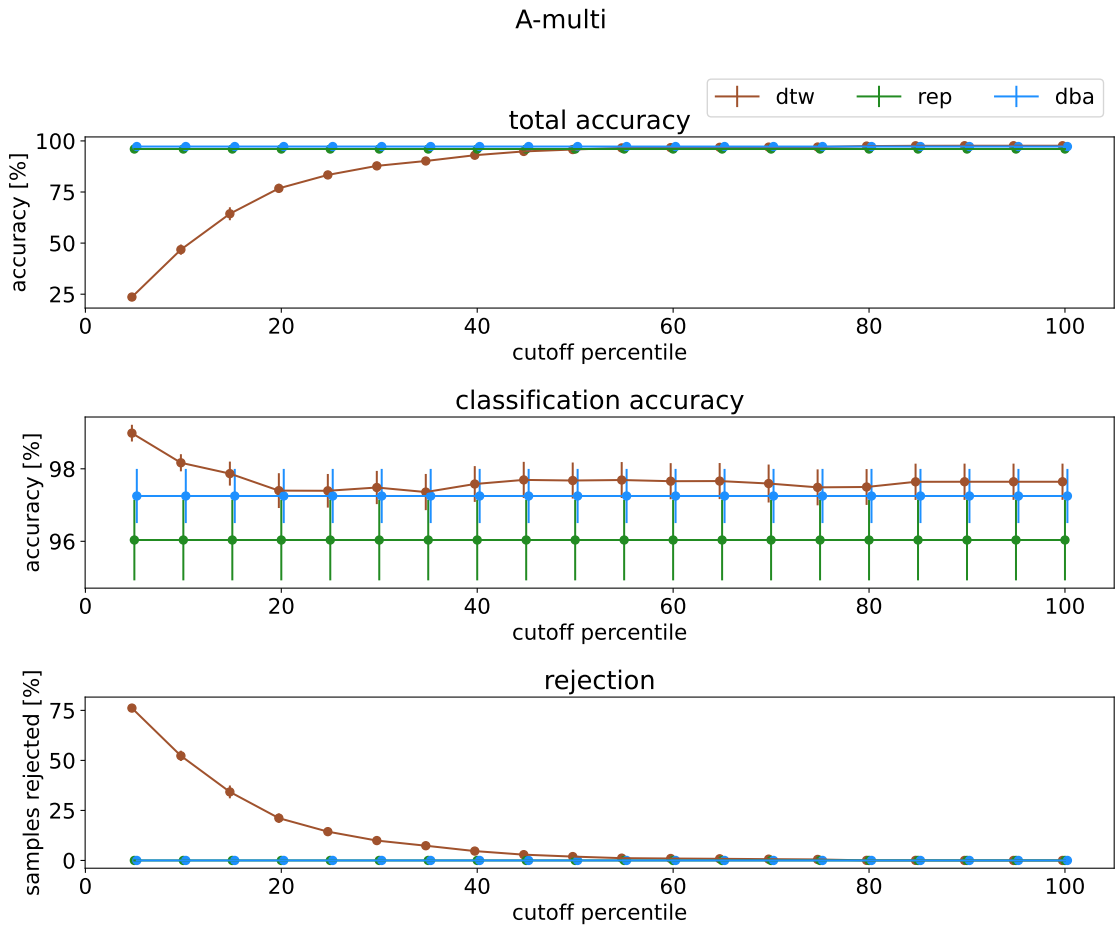

Figure 2.17.: Classification accuracy for different percentiles chosen as cutoff-values for the three classification methods introduced in this thesis for subject A. Classification was performed on the *multiday* dataset with three days serving as training set and the remaining two days serving as test set. This procedure was repeated 10 times with different recombinations. The data-points indicate the average accuracy achieved, the corresponding errorbars denote the standard error on this average value.

which is plausible given the now more heterogeneous dataset due to the fact that different seatings of the sensors are now separated between reference and test set. This is primarily caused by the reduction in overall accuracy by one to two percentage points in accuracy once converged as well as a substantially lower total accuracy reading for low ξ, corresponding to a substantially higher rejection rate, clearly indicating why convergence is slower. If we see consider the number of data-points until convergence, we see about 5 data-points for the merged dataset for both subjects until total accuracy and rejection have converged in the uncondensed case. Taking the same starting accuracy of around 80% as a starting point, we also see convergence within about 5 steps in ξ, indicating consistent behavior between both kinds of splits. The tail for the non-merged dataset for subject B is longer, but the overall slope at the end reaches a realm of diminishing improvements, hence being still approximately consistent with this finding. We do see, however,

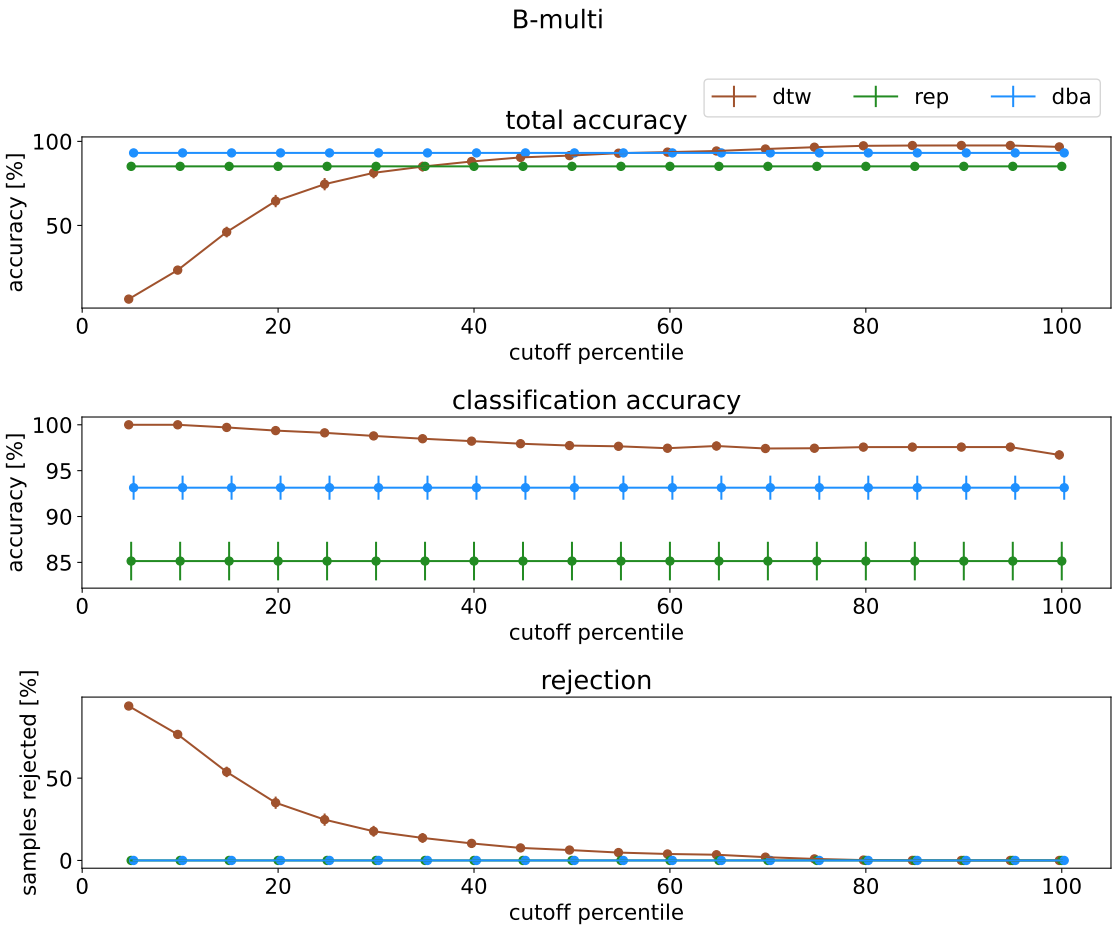

Figure 2.18.: Classification accuracy for different percentiles chosen as cutoff-values for the three classification methods introduced in this thesis for subject B. Classification was performed on the *multiday* dataset with three days serving as training set and the remaining two days serving as test set. This procedure was repeated 10 times with different recombinations. The data-points indicate the average accuracy achieved, the corresponding errorbars denote the standard error on this average value.

a slight reduction in classification accuracy with reduced rejection rates, again continuously for subject B, but only up until $\xi = 20\%$ for subject A, reaching convergence in accuracy there as well. The condensed accuracies, however, point a substantially different picture here, returning constant accuracies with constant standard errors independent of ξ, the strongest convergence result in this analysis and another indicator of the stabilizing properties of condensing. Particularly for subject A we find that the converged uncondensed classification accuracy, while still larger, is mutually within error bounds to the accuracy achieved via barycenter condensing,

Ultimately, this is an excellent result regarding stabilizing classification, as we can confidently establish the 50% percentile of the intracluster distance distribution as a reliable rejection distance ξ for both subjects. This means that we not only improve our classification robustness, but with $\xi = 50\%$ we also have a strong foundation for application of this classification procedure on

a continuous datastream that also contains intervals containing no gesture of interest. A strong boundary ξ will reject these intervals from classification, while still accepting the vast majority of samples that represent one of the gestures in the reference data. Furthermore, due to our strong convergence results especially for the condensing methods (which are preferable not only due to performance advantages but also due to results from previous sections) this boundary can still be chosen substantially tighter, should a user still report incorrect triggering.

2.10. Summarizing the analysis of structural relationships of muscle signatures

We introduced a Dynamic-Time-Warping-based technique to consistently assess the direct similarity of EMG muscle data in the form of EMG signatures for hand gestures with consideration for inconsistent timings of signal characteristics, and have developed a classification system based on this time-corrected similarity measure. We have introduced two optimization setups to improve both performance and robustness while successfully retaining high classification accuracies of the method at reduced computational cost.

We have shown that this technique is not only capable of yielding robust state-of-the-art classification performance, but also allows insight into the structure of gesture sample clusters that form the basis of this outstanding classification performance even under the inclusion of challengingly similar gestures. Consequentially, this also allows the system to be agnostic to the base set of gestures that is used as a foundation. We could show, that our proposed classification method is capable of retaining high accuracies even when exceeding the number of gestures originally supported by the Myo Armband while at the same time, in contrast to the Myo system, not predefining the gestures in question, but leaving this choice up to the user.

The obtained fundamental understanding of the underlying structure also allows for an informed decision in the creation of a gesture set to form the basis for a gesture recognition system.

On this basis, the performance parameters and the set of base gestures, including their number, can be tailored to the specific usage scenario at hand and individualized to a particular user with little effort over the course of a few days or, in case of necessity, even over the course of an afternoon.

We also showed that the accuracy of such a system can be tailored towards the user's preference and use case with regards to precision, i.e. the preference between improving classification accuracy by reducing the risk of misclassification through acceptance boundaries. These boundaries might be chosen more on the side of precision for critical usage scenarios and more liberal and convenient in more lenient situations at the still low but slightly higher risk of mistriggering.

We furthermore demonstrated that muscle signatures are highly individual, making it necessary to tailor such a system specifically to the user. Consequentially this also means that users are free to pick their preferred set of gestures for a particular task, improving acceptance and convenience and potentially reducing the likelihood of mistakes as a consequence of this.

Part II.

Classifying motion with Inertial Measurement Units

In this part we will, for the first time, investigate the question whether motion can serve as a quantitative identifier for aversive, dysfunctional cognitive states in the context of major depressive disorder.

Motion as a physical feature of depression and its severity has also been investigated quantitatively in the literature as early as 1992, where [79] found that patient activity increased after pharmaceutical intervention, already indicating that activity is negatively correlated with severeness of depressive disorder. This is also supported by investigations by [82], who combined self-assessment questionnaires with clinical assessments and the analysis of video footage. Quantifying motor retardation based on that footage, they find both self-assessments as well as professional assessments not fully aligning with motion energy analysis performed afterwards. This shows that motion analysis might serve as an indicator for aspects that can not be estimated by self- or professional assessment via questionnaire. However, this was still an interview (and therefore in a sense laboratory) situation.

In contrast to this, [14] attempted to gather data on mood and cognitive function of depressive patients independent of interviews via daily self-evaluations via Apple watch outside of a standard interview scenario. They complemented the self-evaluations with standard self assessment scales for depression and consequently showed that digital assessment of patient state over wearable technology is equally reliable compared to conventional techniques.

Motor activity with regard to the assessed patient state was investigated by both [72] and [73]. The former used an AdaBoost-based technique to leverage sensor measurements from a wristband and a smartphone to predict depression severeness scores, which are provided by bi-weekly clinician-rated 17-item Hamilton Depression Rating Scale (HDRS-17 or HDRS) [41]. Their available sensor data included a variety of options, including phone usage, communication patterns, sleep characteristics, weather, physiological data and location data, on which not only an HRDS rating was predicted, but also a feature ranking was performed, indicating which kinds of data would best serve as a basis for a reliable prediction of a patients' HRDS score and hence are likely to yield insight into the illness, many of which are related to activity measurements, such as variances in location hinting at increased patient activity. The latter study measured such activity in the form of a wrist-worn IMU on patients diagnosed with major depressive disorder, adjustment disorder and bipolar depression, acquiring said activity data over the course of hospitalization. Leveraging both a bi-level Gaussian Process regression model as well as a Random Forest model, they found that the acquired measurements can be used as an indicator to predict the time of hospital discharge down to an average discharge prediction error of around 25 days towards discharge after around 40 days of admission.

Significant reduced amounts in activity were also found by [68] in patients suffering from late-life depression. Patients included into this study were above the age of 60, not clinically admitted, but were visited three times over the course of the 7-day study at their homes, once at the beginning, delivering the study, once in the middle of the study and once at the end of the study, and assessed on well-being and severity with various methods, among them the Montgomery-Asberg Depression Rating Scale (MADRS) as well as the 15-item Geriatric Depression Scale (GDS-15).

Furthermore, several other studies have attempted to quantitatively assess motion in depressive disorder, focusing, among other indicators, on assessing sleep patterns using activity measurements via wrist-worn actigraphs [11, 13, 46, 64]. More information and further studies can be found in the review by [60], listing 18 different studies attempting quantitative assessments, among other diseases, depressive disorders.

However, to this point all investigations so far have focused on the overall state of motion in the context of major depressive disorder, without assessing its viability to identify acute dysfunctional cognitive states in the context of it. To change this, we will derive an approach that is not

only capable of evaluating motion data with respect to the specific question of differentiating between aversive and non-aversive cognitive states, but is also able to do so autonomously and transparently.

An autonomous derivation of an optimal classification and identification scheme for specific patient data allows to quickly iterate on the development of such a method in cooperation with partners in the clinical and hardware context and also yields a simple extension to other fields of application. It also enables a more dynamic development in terms of process details, such as the focus on the preferred detection of aversive states in contrast to the goal of avoiding overdetection of such states, which may require both a slightly different treatment as well as a different classification configuration for optimal classification performance.

Transparency on the other hand is particularly important for the medical application both in terms of regulations as well as patient acceptance. Furthermore it allows a deeper look into how different components support the resulting classification objective, and an according selection of components most beneficial to this objective.

To derive this procedure, we will first briefly discuss major depressive disorder, which provides the fundamental context of our application, and provide insight into how our method integrates into the MITASSIST project. Afterwards, we will discuss the procedure of recording patient data such that we can identify aversive and non-aversive cognitive states. After outlining the acquisition procedure, we will focus on the individual components that we will use to obtain a data-to-result classification procedure. These are slicing of the time series data, characterization of these slices, both in terms of fundamental data representation as well as data characteristics, transformations of the space of these characterizations to benefit a classification procedure and finally the classification schemes investigated in this thesis as well as a Bayesian procedure to obtain optimal classification scheme configurations.

We will then show that this method is capable of yielding problem-specific optimized classification schemes and investigate the impact of different components onto the resulting classification accuracy. First, we will investigate this with respect to recorded data of individual patients. We will show that the resulting optimal configurations work and what this means with respect to different classification schemes, both in general as well as for each individual dataset considered. We will then investigate the impact of feature space transformations and how they aide the classification objective, finding a strong preference for Linear Discriminant Analysis and similar clustering-techniques. We will also investigate, how these transformations combine with specific classification schemes to achieve the described classification accuracies. Afterwards, we will investigate the relevance of data representations, showing that for aversive states in major depressive disorder, the exact representation of the data is of less relevance.

We will then proceed to investigate how the sensitivity, specificity and precision performance exhibited by different classification schemes on the specific patient-individual datasets allows an insight into the way these algorithms achieve their associated accuracies and its relevance for selecting a configuration for a specific purpose in the context of treatment of major depressive disorder.

We will then repeat and extend this investigation on a dataset that comprises data recorded from multiple patients to see how the acquired data is able to generalize. While we see a noticeable reduction in classification accuracy, we still see that our method is capable of producing advantageous configurations. We again investigate this in terms of general accuracy and the impact of feature space transformations. We will also investigate sensitivity and specificity in more detail both for the mixed as well as individual datasets.

We conclude the part with a proposed optimal configuration of a k-Nearest-Neighbours scheme that is suited to both patient-individual as patient-aggregated data.

3. Major Depressive Disorder

Unipolar depression is one of the most prevalent mental disorders frequently observed and often comorbidities [10] with anxiety disorders as well as alcohol-related illnesses, which together make up more that 35% of all diagnoses in this area. While scientifically based treatment guidelines recommend treatment of medium to severe episodes of depression, also referred to as major depression, the actual resources for psychotherapy are very limited, especially with respect to treatment frequency [27, 87]. Especially the acute phase treatment should provide on demand interventions in real time to address highly stressful faces of rumination and despair.

3.1. Symptomatology

According to the International Classification of Diseases, version 10, (ICD-10, [69]), an unipolar depressive episode is characterized by three *primary symptoms* [10]:

Depressive or subdued mood is typically reported very subjectively and differs between subjects and reporting can range from emphasising negative feelings such as hopelessness or desperation to the reporting of not feeling anything, neither joy nor sadness. In both cases, patients have difficulties to enjoy positive events as well as feel negative events, with high prevalence combined with feelings of anxiety without direct cause, fear of the future and overwhelmed by social interaction. These characteristics are typically reported to be time-stable, with a disproportionally high prevalence occurring in the morning and reduction over the day.

Loss of interest and enjoyment is commonly correlated with a reduction in activity. Due to this reduction, increasing activity again is perceived as challenging to impossible and can typically only be overcome in lighter episodes. Noticeable targets of this reduction of activity are things that can be perceived as exhausting on their own, such as work or housekeeping, but can also affect recreational activities. This symptom is also often referred to as anhedony.

Reduction in drive or motivation and fatigue is a predominantly subjective symptom and often correlated with loss of interest and enjoyment. Whereas the latter shows itself in actual reduction in activity, reduction in drive or motivation and tiredness tends to represent itself as the subjective perception equivalent of it, causing such a reduction in activity as this activity is perceived as exhausting in addition to the lack of joy obtained from it.

In addition to these key symptoms, a larger number of *secondary symptoms* is frequently observed. These include, according to [69] and as described by [10], the following:

Reduced concentration and attention and with it a reduction in both the capability to think and to make decisions is inhibiting the patients capability of staying operational in their daily life and routine, which, to a degree and dependent on the situation, requires constant decision making in varying frequencies. These intervals of attentional focus on excessively processing negative thoughts are referred to as rumination. In most cases patients are unable to end these

intervals by themselves even if instructed in a technique to do so by a therapist. Rumination is accompanied by intensive experience of aversion, anxiety and tension.

Feelings of guilt and worthlessness describe the tendency of patients to doubt their own capabilities in work, social and leisure contexts. These feelings are also reflected and intensified during rumination.

Negative and pessimistic perspectives of the future describe a disproportionally negatively skewed perception of the future or future prospects, including, but not limited to personal development as well as treatment perspectives for the disorder itself.

Suicidality in the form of suicidal thoughts is often reported to be perceived as a way of ending the suffering induced by major depressive disorder. In combination with the previous secondary symptoms regarding negative perceived future perspectives suicide is potentially viewed as the only way out of the patient's condition. Suicidality can either occur in the form of ideation, where suicidal thoughts are contemplated, but not followed up upon, or actual suicidality with a non-zero chance of the patient attempting to end their own life as a perceived last resort. Given the severity of this, the assumption of ideation must obviously not be made lightly. In fact, the reduction in well-being leads to a 30-fold increase in prevalence of suicide among persons affected by depressive disorder [37], with the number of attempted suicides being seven- to twelve-fold higher that the number of completed suicides.

Sleeping disorders are typically observed in combination with major depressive disorder in the form of sleeplessness with regards of falling asleep, staying asleep or waking up early. Tiredness over the day or hypersomnia, a prolonged sleep in the night are rarely reported in combination with major depressive disorder.

Reduced appetite can lead to substantial reductions in food consumption and can be a consequence of a lack of motivation and activity. In severe cases, this can lead to a substantial loss of weight.

The differentiation between unipolar to bipolar depression can be made by contrasting the described negative symptoms with the existence of manic, positive phases. Bipolar depression typically yields phases of enjoyment and positivity, followed by the described symptoms of depression, whereas unipolar depression lacks overly positive phases and predominantly exhibits only the negative aspects. In the project presented here, we will limit ourselves to the inclusion of patients with unipolar depression only.

The severity of major depressive disorder is considered light if at least two of the primary and two of the secondary symptoms are observed, a medium severity is assumed if two primary and three to four secondary symptoms are observed and a severe major depressive disorder is diagnosed if all three primary symptoms and more than four secondary symptoms are observed in conjunction.

There are different courses of an unipolar depression that are described in [10]. The classical depressive episode with full remission describes a singular episode of major depressive disorder that will fully return to normal levels of well-being. Alternatively, a patient can encounter a recurring depression, which describes multiple "single episodes" of major depressive disorder with full remission in between those episodes, but repetition of them. Alternatively, a singular episode with incomplete remission can occur, where the patient only return to previous levels of well-being, but not entirely, with a small reduction in well-being remaining. This condition is

referred to as dysthymia. All of these can also occur in combination, such as a full depressive episode with previous dhystemia. Finally, a depressive episode is considered chronified if there is no noticeable remission for a time span of more than two years.

3.2. Treatment and therapy

As with common physical illnesses, common depression treatments are the reduction of symptoms of Depressive Disorder with the goal of full remission. Correspondingly, the reduction in motivation, focus and attention and loss of capabilities in different extrinsic fields of a patient's life, such as work, leisure or social life, shall be restored as well as the intrinsic, personal well-being, such as motivation, the feeling of self-worth or agency. In case of suicidality, the risk of mortality also qualifies as a direct goal of treatment. In addition to that, the risk of recurrence shall be reduced to retain a continuously stable healthy state without relapse.

To achieve this, we can fundamentally differentiate between three different primary approaches:

Watchful waiting describes a low-profile intervention with the attempt to reduce the effect of symptoms by potentially assistive activities. This includes assisted self-help or technology-assisted approaches that aim to assist the patient in addressing their own situation. This is specifically targeted to low-severity cases of depressive disorder with a chance of remission without substantial intervention and can therefore be a successful approach to aid in the very early or very light stages of such a condition. It is important to emphasise that this does not describe the mere information of a patient about techniques to address their condition, but also assistance in implementation of these. It must, however, be watched closely whether this kind of intervention is appropriate and sufficient to address the specific symptomatology at hand or if other approaches as described below should be advised.

Pharmacotherapy describes the treatment of major depressive disorder by means of medication. Diffentent options exist for this purpose whose approach differs. Given that a pharmacological approach is not the primary target of this work, we will review them briefly and refer to [10] for more detail. Briefly, the classes of substances available for pharmacotherapy include Tri- and Tetracyclic Antidepressants as well as Selective Serotonin or Noradrenaline Reuptake Inhibitors, both of which inhibit the reuptake of Serotonin and Noradrenaline, increasing the central neurotransmission that is operated by these neurotransmitters, or Monoaminoxidase Inhibitors, all of which are classified as antidepressant medication, whereas Lithium and Phytotherapeutics are not classified as such, but the mood stabilization effects of the former can also contribute to an increased well-being. Unfortunately, the exact effect mechanisms of antidepressant medication is still not entirely understood.

Psychotherapy describes, as defined by [90], the treatment of the basis of an intervention with predominantly psychological means. Review suggests the effectiveness of psychotherapy as on-par with a pharmacological treatment with anti-depressants [17, 29, 45]. These studies were predominantly performed with ambulant and not with hospitalized patients, so the scientific basis for inpatient treatment is weaker, but the existence of evidence for ambulant treatment allows the assumption of a noticeable degree of effectiveness for inpatient treatment as well.

Five different factors of effect for psychotherapy are described in the literature [10]:

Firstly, the *therapeutical relationship to the patient* allows, by establishing a qualitative and systemic relationship with the patient and valuing their worries and feelings, for the most significant effects in reducing said worry and negative feelings as well as reducing impact of the aforementioned symptoms [12], independent of the specific specialized variant of psychotherapy.

Further effects that are described are *resource activation*, which describes the individual traits and characteristics a patient is contributing to the therapeutic effort, in particular still existing motivation that can be activated this way. *Problem actualization* describes the presentification of difficulties that have lead to the patient's condition in the first place. By making these problems more visible to the patient, the patient is enabled to realize them and regain agency over them, which is described as part of *problem resolution*, where the patient is supported in the development of strategies to address and potentially correct these specific problems with the goal of removing the underlying root cause for the patient's condition. Finally, *motivational clarification* builds on this by enabling the patient to reflectively realize the underlying root causes and enabling the patient to autonomously address these problems.

Relevant for the effectiveness of this that the patient gains the experience to be understood and sufficiently allowed to communicate their current view on the problems or what they perceive as such, including, but not limited to, their understanding of the situation, root causes and background. From that, different approaches for psychotherapy can be attempted. These include *behavioral therapy*, which focuses on aspects of learned helplessness and negative reinforcement, *cognitive therapy* which focuses on the existence of a negative cognitive spiral of continued worry, *psychodynamic therapies*, which focuses on potentially traumatic or impairing interpersonal experiences that could not be processed accordingly in the past as well as feelings of loss and humiliation that are potentially incorrectly perceived as the patient's fault. A fourth kind of therapy is *systemic therapy*, which focuses on the social context and the environments as root causes for the patient's condition. A solitary psychotherapy program specifically tailored for depression treatment is *interpersonal psychotherapy*, which focuses on overcoming of psychosocial root causes, such as unaddressed grievance, role conflicts or social conflicts. *Person-centered therapy* finally attempts to assist the patient in the resolution between current and idealized self-image, which can lead to the aforementioned symptoms. Effectiveness of psychotherapy have shown to be equal to pharmacotherapy, albeit for severe cases conditions the effect latency can be longer as for pharmacotherapy (see also [92]). After the latency period, however, both methods have shown equal superiority to a placebo therapy [18]. Also, several studies indicate the effectiveness of psychotherapy already after short periods of three to five weeks of time [35,39,40] and to be superior in their effect of relapse prevention after discontinuation of treatment [17].

Combination of Pharmaco- and Psychotherapy is in a sense a fourth way of treating major depressive disorder and is particularly useful for more severe cases, harbouring the positive effects of both methodologies. Also, it has been shown that the combination of pharmacotherapy with a psychotherapeutical approach increases patient compliance with regards of taking their prescribed medication. [16, 40, 44, 65].
Beyond that further therapy options are electroconvulsive, sleep deprivation or light therapies as well as physical activity, repetitive transcraniel magnet stimulation or vagus-nerve-stimulation. For more details on these, refer to [10].

The official national treatment guidelines [10] recommend a low-profile intervention or psychotherapeutical approach for light severities, either a pharmaco- or psychotherapy for cases of medium severity and a combination of pharmaco- and psychotherapy for severe cases of Unipolar Depressive Disorder.

3.3. The Mitassist project

These aforementioned treatment options are known to yield an improvement to the patient's condition and well-being. However, in contrast to physiological conditions, major depressive

disorder is strongly correlated with impacts on the patient's capability to perform basic tasks and their personal agency, which is significantly reflected in the common inability of patients to control symptoms like rumination at the actual moment of their occurrence in the patients daily life even if instructed how to do this by a therapist. As a consequence, the experience of helplessness is reinforced due to the absence of therapeutic support outside the therapy sessions. Therefore, therapeutical success it strongly dependent on the patient's capability to adhere to the therapeutical means and goals. If this cannot be guaranteed, hospitalization is potentially a consequence, which, however, has a noticeable impact on the patient's life as they are drawn from their normal life. On the other hand, keeping the patient in their normal routine at home might leave them unsupervised and unassisted to their condition, which also can impede the progress of therapy, particularly when daily structure is already deteriorated.

To bridge this gap, the Mitassist project, an abbreviation for **M**ulti**I**mmersive **T**herapy **Assist**ance proposes a novel approach to supplement patient treatment by monitoring and targeting symptoms via automated recording of motion data that can be used to monitor the patients well-being. In particular, the wearable device proposed by the Mitassist project contains sensors that can be used for drug level monitoring, monitoring a patient's heart rate as well as motion data in the form of IMUs.

If this data can be used to detect dysfunctional states of patients diagnosed with depressive disorder, ambulant treatment can be monitored and supported by real-time interventions externally, allowing for patients to be treated at home, which can increase well-being as well as improve recovery and alleviate the strain on the healthcare system that would otherwise need to hospitalize these kinds of patients. Furthermore, such a system can be used to quantify a patient's recovery trajectory, allowing quantitative insight whether therapeutic options have an effect and potentially allowing automated intervention. This would allow a level of supervision that would typically only be achieved in an inpatient context.

A wearable system like the MITASSIST setup would allow tailored interventions integrated in the patients daily routines, without impacting the patient's daily life in a way that a hospitalization would, allowing the patient's social environment to remain a contributing factor to recovery as well as removing strain from the healthcare system by reducing the number of patients that remain without effective treatment for a long time, a situation that is already common in the German healthcare system.

This, however, requires the derivation of well-performing classification procedures that are able to assess the difference between dysfunctional and non-dysfunctional cognitive states of the wearer. Therefore, we will in this part derive a procedure that allows to find an optimized classification procedure for such a problem given labeled motion data. To do so we focus on classifying intervals in the daily life of depressive patients with and without rumination in order to identify this aversive state as a target for automated assistance by a wearable like e.g. offering attentional shifts.

4. Assessment of aversive mental states in the daily life of patients with mental disorders

The aforementioned diagnostic assessment investigates the overall trend in motor activity, while patients of depressive disorders report that even over the daily course of the depressive syndrome, their level of well-being does vary significantly between less pronounced expression and severe symptoms. Regarding ruminations, patients treated with conversational psychotherapy either in addition or instead of pharmaceutical treatment are instructed to observe the occurrence of the symptom and to interrupt it. However, as these kind of symptoms are part of the disorder, it proves difficult for patients to escape the symptomatic cycle of acute depressive symptoms. An external detection of acute expression of cognitive symptoms like ruminations could therefore be a beneficial part of treating depressive disorder, allowing for defined intervention. To base an intervention on such a kind of recognition system, we of course need to make sure that it is both capable of differentiating between functional and dysfunctional states, as well as allowing an easy understanding of how a classification result is derived, both for regulatory reasons as well as actually providing explanatory power instead of just a result, as also remarked by [96].

Hence, we aim at investigating means of optimally classifying obtained patient data to maximize both explanatory power as well as obtaining means of classification with high accuracy. To do so, we will first define the underlying acquisition hardware, the application and procedure to acquire the current cognitive state of patients as labels as well as the process of obtaining the data foundational to an attempt to seek optimal recognition of dysfunctional cognitive states. We will then outline a step-by-step procedure expanding our data into a space of classification approaches that covers a variety of different attempts to classify such data, to then formulate the bi-level optimization problem that will yield an optimized classification procedure to capture the current mental well-being of a patient based on measurements automatically acquired.

4.1. Data acquisition: measuring patients within their daily routine

Our analysis will be based on hardware that is specifically developed and tailored towards the aforementioned goal of classifying patient well-being during the course of their illness. The data will be acquired leveraging a wearable device that is equipped with two inertial measurement units (IMU), that each yield three-axis data on linear acceleration as well as three-axis angular velocity. One of the sensors is located at the wrist, the other is located directly at the elbow. Furthermore, the device is equipped with a light intensity senor mounted against inside of the forearm close to the elbow, which is intended to further complement the following analysis with heart rate variability data, which is not yet included into the current analysis. A depiction of this device equipped to a human being can be found in fig. 4.1.

In addition to the recording of motion data, patients are equipped with a smartphone that assesses their mental state and well-being on an hourly basis as long as the wearable measurement

Figure 4.1.: The demonstrational recording device developed within the Mitassist project to acquire motion data of patients during their normal daily routine in worn form on a human being.

device is equipped and paired with the supplied phone. Patients have the option to self-assess their well-being, differentiating between one non-dysfunctional state as well as rumination and sorrow as two dysfunctional states and how long they have been in this state. This yields an annotation of each data-point with either a dysfunctional, a non-dysfunctional as well as an unknown label that serve as the foundation of the following classification of motion data. Said states had been defined for the patients by therapeutic personnel beforehand as part of the treatment procedure and are assumed to be known and understood by the patients included into the study. After instructions of participants, the study was conducted as an unsupervised, discontinuous data acquisition over the course of ideally one week. Not all participants participated for the full week, if study participation was terminated before that, but consent for data usage remained, the data was retained for research purposes. Participants that were considered eligible for participation in this study if they fulfilled the following list of *inclusion criteria*:

- male or female stationary patients with a diagnosed major depressive disorder, i.e. either the first depressive episode (ICD-10-code F32) or a recurring depressive episode (ICD-10-code F33) with at least medium severity
- 18 years of age and above
- written consent after detailed information
- sufficient competences with a Smartphone

Participants were, even if fulfilling the above criteria, not included in the study if they met at least one of the following *exclusion criteria*:

- absence of capability of being informed or consenting
- other mental illnesses occurring at the same time, such as addictions, bipolar disorder, mental disorders of ICD-10 class F2x.x, such as Schizophrenia, organic mental disorder, neurological illnesses, illnesses of ICD-10 type F0x.x, such as Alzheimer's disease or dementia, or ICD-10 type F1x.x, which contains mental and behavioural disorders due to substance use
- acute suicidality

- other severe physiological diseases
- limited mobility
- consumption of illegal drugs in the last month
- diagnosed alcohol addiction without abstinence in the last 2 months
- consumption of cannabis in the last two weeks
- addictions, either acute or anamnestic
- pregnancy, nursing period or positive pregnancy test

The study was approved by the ethics board of the Göttingen Medical University Center and the center's data protection officer. Patients were recruited at the Asklepios Medical Center Göttingen. Datasets were quasi-anonymized including only such demographic data in the study dataset that would not allow identification of an individual participant without a study ID-table. As a result, data of 29 participants could be acquired. However, due to not all participants conducting the full length of the study and also participants identifying intervals of different lengths for dysfunctional and non-dysfunctional mental states, not individual acquisitions yielded equal amounts of data. This has consequences for the individual studies conducted in this thesis as not all participants could meet the required investigation criteria. The specific selection criteria for the specific investigations in this thesis are listed with the according investigations.

5. Data preprocessing and classification methods

Given the acquired data, we will now derive an optimization procedure that is tailored towards yielding best possible classification results while still retaining explanatory power and understanding how these results were derived.

5.1. Slicing temporal data

As the final goal of our research aims at live-detection of dysfunctional depressive states, we first need to decide how to setup our proposed procedure, such that live-analysis and live-detection is possible. Our underlying data consists of data time series, i.e. measurements over time. To identify events on such time series, some techniques operate on the entire series, such as Dynamic Time Warping based techniques, Derivative Transform Distance [31] or Complex-Invariant Distance [4]. However, as the data is accumulating over time, the amount of data we can employ analysis on continues to increase monotonically, however, time to far into the past is unlikely to yield insight into the current state of the patient. To deal with this, many techniques common in time series analysis use a window approach, such as for example by Symbolic Fourier Approximation-based techniques [84, 85], shapelet based approaches [32–34, 48, 101], also in an unsupervised fashion [102], or forest-based techniques, either via shapelet [54] or proximity [63] approaches. To control the amount of past data that is considered, we also employ a window-slicing approach, subdividing the data into segments of defined length. This allows for a precise control of the amount of data leveraged for classification, and as a direct consequence, defines a maximum reaction delay for the classification to trigger if a dysfunctional event is detected. The label of such a single time slice will be assigned based on the label that is most often assigned to data-points within this data slice. In case of a tie, the second half of the window will take precedence over the first, prioritizing more recent data. Also, a number of deep-learning techniques has been proposed, see [52] for a review.

5.2. Characterizing time slices

While the aforementioned methods of time series analysis are tailored towards finding patterns in time series directly, the goal of this approach is slightly, but noticeably different: In contrast to the previous part, we are now no longer seeking direct patterns in the data, but instead look for abstractions that allow us to quantify a slice of data in a way that allows to differentiate between dysfunctional and non-dysfunctional mental states. As we are not searching for patterns in the actual shape of the time series, but in the information it represents, we will follow the example of, among others, [**?**] and [73], we will employ classical pattern recognition techniques. To properly work, these methods rely on comparable inputs for comparable samples, in particular the ordering of information in the input to these techniques must be constant to ensure proper comparability. To ensure this, we will abstract each time slice into a characterization vector that contains a set of properties of the underlying time slice.

In the absence of prior work in discriminating between levels of patient well-being, it is unclear which kind of information best represents either of the two states taken into consideration. Therefore we will include several representations of data in the set of possibilities to optimize for the best representation. As outlined in section 4.1, we directly obtain both linear accelerations $\ddot{x}$ as well as angular velocities ω from the study setup. Based on this, three more properties are derived: First, we will integrate the linear accelerations twice over the window width to obtain a window of equal length that contains for every data-point T_i both the pseudo-velocity

$$\dot{x}_i = \int_0^{T_i} \ddot{x} dt \tag{5.1}$$

as well as the pseudo-positions

$$x_i = \int_0^{T_i} \dot{x} dt \ . \tag{5.2}$$

We consider them as pseudo-velocities and pseudo-positions respectively, because the correct estimation of $\dot{x}$ and x would require the base positions x_0 and velocities v_0 as integration constants. Obtaining these reliably, however, requires both drift correction as well as initial calibration of the device to find the initial position and velocity. Drift in the context is a consequence of the base measurement error every measurement device is subject to. Every measurement $\ddot{x_s}$ acquired of the real value of $\ddot{x}$ will in reality contain two types of errors: noise σ and a base offset ϵ:

$$\ddot{x_s} = \epsilon + \sigma \ . \tag{5.3}$$

While the measurement noise can typically be modeled by a zero-mean distribution, i.e.

$$\int_0^T \sigma dt \approx 0 \ , \tag{5.4}$$

and hence cause a constant magnitude of error over time, the base offset describes a fundamental bias in the sensor that is not zero-centered, hence monotonically increasing over time, resulting it drift and consequentially an increasingly incorrect assessment of x and $\dot{x}$, as both are estimates over integrations since the beginning of the measurement. To avoid working on data that must be assumed to having become increasingly incorrect over time, we will only consider pseudo-positions and pseudo-velocities, taking this problem into consideration explicitly. Consequentially, as we will not attempt to control drift, especially given that we do not have control of the experimental environment, as the measurement setup is worn by patients during their daily routine without supervision, we also refrain from introducing an explicit sensor calibration procedure, as this would complicate the operation of the measurement setup for the patient while yielding little usable information for us, as the sensor system will drift out of that calibration very soon. Furthermore, we will integrate the angular velocities over the window width to an orientation off a default quaternion q_0, i.e.

$$q_i = \int_0^{T_i} \frac{1}{2} q_0 \boldsymbol{\omega} dt \tag{5.5}$$

under the same assumptions that we outlined above for x and $\dot{x}$. This way, while we will not be able to asses actual physical orientation, velocity and position, we will be able to investigate if these kinds of representations are more indicative for the state of patient-well-being that we attempt to detect.

We furthermore will compute the Fast Fourier Transform of every window, yielding an additional time series for every series discussed so far, to see if periodicities in the measurements might be indicative for the assessment of patient-well-being. To avoid classification instabilities due to slight shifts in the detected frequencies in Fourier space, we will divide the resulting representation in Fourier space into groups, which are integrated over. Assessing the data in Fourier space is then done based on these groups.

To now ensure that each time slice will yield a comparable vector characterization, we will now remove the time component from the window entirely, instead computing characteristics on every time slice s that will be used to fill locations in the resulting characterization vector. The representation candidates we use are as follows:

Minimum The minimum value over the entire slice, i.e. $min(s)$.

Maximum The maximum value over the entire slice, i.e. $max(s)$.

Mean The mean value over the entire slice, i.e. $\bar{s} = \frac{1}{I}\sum_{i=1}^{I} s_i$, also known as the first statistical moment.

Variance The variance over the data slice, also known as the second statistical moment:

$$\sigma^2 = \frac{1}{I}\sum_{i=1}^{I}(s_i - \bar{s})^2 \tag{5.6}$$

Standard deviation The standard deviation, defined as $\sigma = \sqrt{\sigma^2}$.

Skewness The skewness of a distribution, defined as

$$\mu_3 = \frac{1}{I}\sum_{i=1}^{I}(s_i - \bar{s})^3 \ , \tag{5.7}$$

and illustrated in Fig. 5.1.

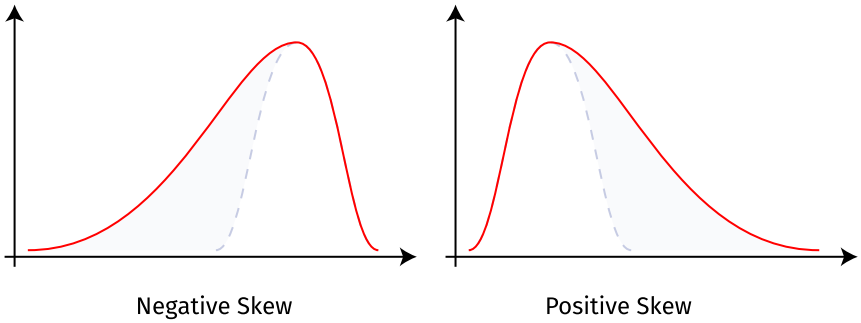

Figure 5.1.: Illustrative skewed distribution, with both positive and negative skew [42].

Excess kurtosis The excess kurtosis of a distribution of data compared to a standard normal distribution (which has a kurtosis of 3), defined as

$$\mu_4 = \frac{1}{I}\sum_{i=1}^{I}(s_i - \bar{s})^4 - 3 \ , \tag{5.8}$$

and illustrated in Fig. 5.2. For reasons of brevity and compactness, it will be referred to in the following also simply as kurtosis, especially in figures.

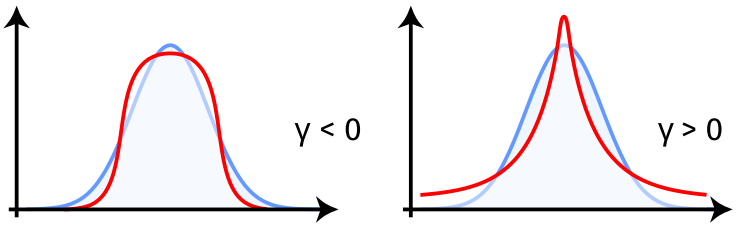

Figure 5.2.: Positive and negative excess kurtosis (red) in comparison with a reference distribution (blue) [1].

Amplitude The amplitude of the signal over the time slice s, i.e. $max(s) - min(s)$.

Relative amplitude The quotient between the signal limits instead of their difference, i.e. $\frac{max(s)}{min(s)}$.

Energy We take inspiration from the common description of energy of classical physical systems, such as kinetic energy ($E_{kin} = \frac{1}{2}mv^2$ with mass m and velocity v), rotational energy ($E_{rot} = \frac{1}{2}I\omega^2$ with inertial moment I and angular velocity ω), potential energy of a spring ($E_{spring} = \frac{1}{2}kx^2$ with spring constant k and displacement x), energy in an electric inductor, such as a coil ($E_{ind} = \frac{1}{2}LI^2$ with inductance L and current I) or energy in an capacitative electric field ($E_{cap} = \frac{1}{2}CU^2$, with capacity C and voltage U), to derive an estimate akin to energy for our data signals. Using the structure

$$E = \frac{1}{2}cs^2 \tag{5.9}$$

with a capacity constant c and a signal value s. As the physical measure for capacity depends on the type of signal used, we leverage the fact that every attempt of classification based on reference data is in itself a comparative analysis to generally define

$$c = 2 \ . \tag{5.10}$$

This way, we retain independence of this measure of specific types of input data, making processing much more flexible and at the same time simplify our energy estimate to

$$E = s^2 \ , \tag{5.11}$$

eliminating an unnecessary division in the process.

For some properties that we do use later on, such as velocity and angular velocity, we actually obtain the mentioned energy estimates, namely E_{kin} and E_{rot}, just renormalized to a default capacity measure for each.

Furthermore, we compute the aforementioned properties on subdivisions of the underlying data window, to preserve some temporal information while still removing the the time axis as such. The optimizer is free to combine properties and data sources as it sees fit.

5.3. Optimizing for separability

Now that we have established characterizations of the underlying data, we could proceed to use these characterization vectors as input vectors to an assortment of classification algorithms. However, we can still encounter a set of difficulties: Firstly, depending on the datasets and characterizations used for the input vector, this vector can grow to a significant size, and each vector component represents one dimension in the resulting dimensionality space, possibly making the detection of patterns and clusters in that space cumbersome. Secondly, the resulting space might be suboptimally distributed to serve as a basis for the application of classification. To address both issues, we will extend the aforementioned procedure by additionally transforming the space of resulting characterization vectors. We include several transformation options into the analysis, some of them are retaining the number of dimensions of the original space while others reduce dimensionality. These transformations t can be applied to a dataset $D = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}) \in \mathbb{R}^{n \times m}$ with corresponding labels $(y_1, \cdots, y_n)$, to obtain a different distribution of samples in D. The set of non-reductive transformations, that retain space dimensionality, contains the following mappings:

Linear The Linear transformation will not transform the data in any way, i.e.

$$t(\boldsymbol{x}) \mapsto \boldsymbol{x} \ . \tag{5.12}$$

Cosinus (Cos) Using a cosinus transformation will transform the data as the cosinus of its values, i.e.

$$t(\boldsymbol{x}) \mapsto (cos(x_1), \cdots, cos(x_n))^T \ . \tag{5.13}$$

We include this to explicitly hint towards possible periodicities in the dataset.

Scale The scale transformation rescales the data to zero-mean and unit-variance, i.e.

$$t(\boldsymbol{x}) \mapsto \boldsymbol{x'} \tag{5.14}$$

$$\text{s.t. } \frac{1}{n} \sum_{i=1}^{n} {x'}_{i,m} = 0 \ \forall \ m \tag{5.15}$$

$$\frac{1}{n} \sum_{i=1}^{n} {x'}^2_{i,m} = 0 \ \forall \ m \ . \tag{5.16}$$

RobustScale (Robust) Similar to the standard scaling mentioned above, this transformation will rescale not according to mean and variance, but according to the 25th, 50th and 75th percentile, such that

$$p50(D) = 0 \tag{5.17}$$

$$p25(D) - p75(D) = 1 \ . \tag{5.18}$$

By doing so, a similar rescaling like with mean and variance is achieved, however, this method is more robust against outliers.

In addition to these transformations, we also employ dimensionality-reducing transformations. These transformations are set up to, if applied, retain at least 2 and at most 20 dimensions of feature space. Furthermore, for reasons of implementation, reductive transforms will always reduce the dimensionality d of a feature space by one, even if $2 < d < 20$. The set of reductive transforms consists of the following mappings:

Principal Component Analysis (PCA) A principal component analysis attempts to transform a given dataset D into a new ortholinear coordinate frame set by the eigenvectors of the data's covariance matrix Σ_D, which maximizes the variance of the data each dimension. In other words, be Σ a square matrix with real-valued, linearly independent entries, then there exists a transformation U, such that

$$\Sigma = U\Theta U^{-1} \tag{5.19}$$

with Θ being a diagonal matrix and U being the transformation from the original space Σ was computed in to its eigenvector space. Consequentially, samples are then transformed as

$$\boldsymbol{x} \mapsto U\boldsymbol{x} \ . \tag{5.20}$$

In addition to maximizing the variance, it also removes the correlation of the different components by nature of transforming into the eigenvector space. By constraining the number of dimensions used of the transformed space to the k ones with highest variance, a problem's dimensionality can also be reduced to k dimensions, alleviating difficulties for high-dimensionality problems both in method stability as well as computational complexity.

Singular Value Decomposition (SVD) Similar to PCA, a Singular Value Decomposition will decompose a matrix $C \in \mathbb{R}^{n\times m}$, such that

$$C = U\Sigma V^T \tag{5.21}$$
$$\text{s.t. } U \in \mathbb{R}^{n\times n} \tag{5.22}$$
$$\Sigma \in \mathbb{R}^{n\times m} \tag{5.23}$$
$$V \in \mathbb{R}^{m\times m} \tag{5.24}$$

with Σ being a diagonal matrix of effective rank $min(n, m)$. We then define $D_{SVD} = U\Sigma$ as the new representation of the dataset C. This decomposition serves as a transformation while preserving full rank of the dataset. Additional samples can then be transformed accordingly via

$$\boldsymbol{x} \mapsto \boldsymbol{x}V^T \ . \tag{5.25}$$

We will now extend this to a *truncated SVD*, which will allow reduce the number of active features within the dataset by obtaining a matrix C_k that has a reduced rank $k < rank(C)$ by finding a rank-reduced Σ_k, such that

$$\min_{C_k = U\Sigma_k V^T} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} \left(C_{ij} - C_{k,ij}\right)^2} \ . \tag{5.26}$$

The resulting transformation rule to obtain C_k then needs to be applied to new or tested samples as well prior to the aforementioned transformation of the data. [88]

Independent Component Analysis (ICA) Independent Component Analysis attempts find a representation for a data sample $\boldsymbol{x} \in \mathbb{R}^n$ in form of a linear combination of basis functions s_i, such that

$$\boldsymbol{x} = \sum_{i=1}^{n} a_i s_i = A\boldsymbol{s} \ . \tag{5.27}$$

This does allow for an alternative representation of a given dataset. To leverage this in practice, however, we need to find a set $\{s_i\}$ of independent components to choose as basis functions. The underlying assumption we need to make is that these basis functions must not be of Gaussian nature, as the sum of Gaussian distributions, as it would be the case in eq. (5.27), will itself result in a Gaussian distribution with a modified mean, i.e. for two uncorrelated points (with, for simplicity) zero mean and unit variance) we find

$$p(s_1, s_2) = p(s_1)p(s_2) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{s_1^2}{2}\right) \frac{1}{\sqrt{2\pi}} exp\left(-\frac{s_2^2}{2}\right) = \frac{1}{2\pi} exp\left(-\frac{s_1^2 + s_2^2}{2}\right) \ , \tag{5.28}$$

which itself is a Gaussian distribution and hence identically distributed with just a different mean and variance, making it impossible to retrieve the distributions $p(s_1)$ and $p(s_2)$ from it independently, which we need to do in order to find a suitable set of basis functions [51]. If we are merely speaking about the sum of these basis functions, we also find that the Central Limit Theorem states that by summing over multiple Gaussian distributed samples we will obtain another Gaussian distribution, just with a smaller variance. So, to reliably estimate a set of basis functions, they need to be chosen very non-Gaussian. A common estimate for non-Gaussianity is a distribution X's kurtosis [51], a distribution's fourth statistical moment, i.e.

$$\mu_4(x) = E[x^4] - 3\left(E[x^2]\right)^2 \ , \tag{5.29}$$

which provides a number of convenient features, such as linearity, which can easily shown using the above definition. It also follows the definition of the kurtosis by Fisher, which is constructed such that Gaussian distributions have a vanishing kurtosis. As we want to obtain basis functions that are as non-Gaussian as possible, we can now reformulate the computation procedure to obtain the independent components. We first invert eq. (5.27) to

$$\boldsymbol{s} = A^{-1}\boldsymbol{x} = W\boldsymbol{x} \tag{5.30}$$

to obtain a system of equations defining our basis set. We can then formulate the requirement onto our inverted weight matrix W to be as non-Gaussian as possible given our data as

$$W^* = \underset{W}{argmax}\ \mu_4(W\boldsymbol{x}) \ , \tag{5.31}$$

yielding a set of optimally suitable basis functions s_i as a corollary. For a more detailed derivation, see [51]. By choice of the number of basis functions employed, an ICA can also be leveraged to achieve a reduction in problem dimensionality.

In practice, an additional common step performed in preparation for ICA is whitening, which describes the removal of cross-correlation and non-unit variance in the input data, i.e. transforming the data, such that its covariance matrix equals the identity matrix. See also [51] for a detailed discussion of this, as well as the FastICA algorithm that is practically used to efficiently execute an ICA.

IsoMap (IM) The IsoMap, proposed by [91], attempts to find a dimensionality-reduced representation of a given dataset while preserving its possibly non-linear structures, preserving

inter-element distances in particular, making it an extension of Multidimensional Scaling. While Multidimensional Scaling will estimate a point-wise geometric distance between all available samples, which will ignore nonlinearly shaped distributions of data within the original data space, IsoMap leverages a neighbourhood-graph G to establish distances between samples. To create this neighbourhood graph, samples are considered as graph nodes of G, connected to nodes in close proximity. This proximity can either be established by considering every neighbouring node within a distance ϵ, or alternatively, a node will connect to its k closest neighbours, with the edges being weighted by their distance $d(x_i, x_j)$. We can then find the shortest path $(x_{i,j,1}, \cdots, x_{i,j,L})$ between samples x_i and x_j and establish a graph distance between those two samples via

$$d_G(x_i, x_j) = \sum_{l=1}^{L-1} d(x_{i,j,l}, x_{i,j,l+1}) \tag{5.32}$$

which is the sum of all edges on the shortest path between x_i and x_j. Consequentially, we can establish a distance matrix

$$M_{D,i,j} = d_G(x_i, x_j) \ , \tag{5.33}$$

which, compared to the naive distance matrix $M_{i,j} = d(x_i, x_j)$, better captures structurally bound distributions of points, for example in structures such as the "Swiss Roll", see [91] for illustrations. We furthermore define an operator $\tau(M)$, such that

$$\tau(M)_{ij} = -\frac{\left(\delta_{ij} - \frac{1}{N}\right) M_i j^2 \left(\delta_{ij} - \frac{1}{N}\right)}{2} \tag{5.34}$$

which converts distances to inner products [91] with $\left(\delta_i j - \frac{1}{N}\right)$ describing the "centering matrix" [**?**]. This then allows us to obtain a representation of the data in a reduced dimensionality Y by optimizing

$$\min_{D_Y} \left\| \tau(D_G) - \tau(D_Y) \right\|_2 \ , \tag{5.35}$$

i.e. finding the distances between the samples in the lower dimensionality space such that their distances optimally resemble the distance distribution in the original space, resulting in a dimensionality-reduced dataset.

The global optimum of eq. (5.35) is found by setting the coordinates $y_i \in Y$ to the top m eigenvectors of $\tau(D_G)$ [91].

Locally Linear Embedding (LLE) Locally Linear Embedding attempts to find a representation of high dimensional data in a lower dimensional space by directly constructing a reconstruction of a data-point in lower dimensional space based on its neighbours, introduced by [83]. Similar to IsoMap, the neighbours to be considered can either be determined by a maximal distance ϵ or alternatively by a number of k nearest neighbours considered. Once the neighbours x_j considered for a sample x_i are established, we can find a set of weights w_{ij}, such that x_i can be optimally reconstructed by them, i.e. we want to find w_{ij}, such that we minimize the overall reconstruction error:

$$\min_{w_{ij}} \sum_{i=0}^{N} \left\| x_i - \sum_j w_{ij} x_j \right\|_2 \ , \tag{5.36}$$

linearly reconstructing each point based on its surrounding neighbours. After establishing the reconstruction weights in original space, we can then pick samples y_i from a space of lower dimension, such that the reconstruction weights found in (5.36) apply to this space as well, i.e.

we want to find y_i, such that

$$\min_{y_i} \sum_{i=0}^{N} \left\| y_i - \sum_j w_{ij} y_j \right\|_2 \tag{5.37}$$

to obtain an optimal representation of the original data in a lower dimensional space Y. The optimal embedding for this problem is found by diagonalizing

$$M_{ij} = \delta_{ij} - w_{ij} - w_{ji} + \sum_k w_{ki} w_{kj} \tag{5.38}$$

and using M's bottom $m+1$ eigenvectors [47, 83].

Spectral Embedding (SE) Spectral Embedding, as proposed by [5], is designed to represent data samples by in a lower dimensional space based on the Laplacian of the adjacency graph. To obtain this Laplacian, we will first have to obtain this underlying adjacency graph G, which is achieved the same way it is achieved in case of Locally Linear Embedding and IsoMap. We furthermore raise the number of neighbours considered to 10% of the sample size to obtain a more dense graph representation to compute the Laplacian of, with each sample representing one node of the graph connected to all considered nearest neighbours. The edge weights between connected nodes are computed as the euclidean distance $d(x_1, x_2) = \|x_1 - x_2\|_2$between the two involved node samples, corresponding in a weight matrix

$$W_{ij} := \begin{cases} d(x_i, x_j) & \text{if } x_i, x_j \text{ connected} \\ 0 & \text{otherwise} \end{cases} \tag{5.39}$$

We then define the diagonal weight matrix

$$D_{ii} = \sum_j W_{ij} \tag{5.40}$$

and based on this the Laplacian matrix

$$L = D - W \tag{5.41}$$

to then solve the eigenvector problem

$$L\boldsymbol{f} = \lambda D\boldsymbol{f} \ . \tag{5.42}$$

Then with, without loss of generality, $\boldsymbol{f}_0, \cdots, \boldsymbol{f}_{n-1}$ solutions to the above eigenvector problem ordered by ascending eigenvector, we can define the subspace constructed by $\boldsymbol{f}_1, \cdots, \boldsymbol{f}_k$ to be the k-dimensional spectral embedding of the original sample space [5].

Sparse Random Projection (RND) The general idea of random projections in general is, given a sample $\boldsymbol{x}$ in a sufficient high dimensional space, such a sample can be linearly mapped into a lower dimensional space as $\boldsymbol{y}$ while approximately preserving the space's structure, i.e. retaining relative distances between corresponding samples. If this holds true, this projection can be randomly chosen, as long as a sufficient amount of dimensions is retained to guarantee approximate preservation of relative distances of samples. The amount of dimensions required to achieve this can be deduced using the Johnson-Lindenstrauss lemma [53], which states

Theorem 3. *Be $0 < \epsilon < 1$, be furthermore $X \in \mathbb{R}^N$ a set of m points and be $k > \frac{8ln(m)}{\epsilon^2}$. Then there exists a linear mapping*

$$M : \mathbb{R}^N \to \mathbb{R}^k \tag{5.43}$$

$$\boldsymbol{x} \mapsto \boldsymbol{y} = M(\boldsymbol{x}) \tag{5.44}$$

such that

$$(1-\epsilon)\left\|x_i - x_j\right\|^2 \leq \left\|M(x_i) - M(x_j)\right\|^2 \leq (1+\epsilon)\left\|x_i - x_j\right\|^2 \ . \tag{5.45}$$

This confines the number of dimensions k of the random subspace given a measure of accuracy ϵ in preserving the original space's structure.

After deciding for a dimension for the random subspace, we then need to pick a random linear transformation. Following the sparse random projection approach proposed by [2] with a modified matrix population rule [62, 71], we can construct a random linear transformation, such that

$$M_{ij} := \begin{cases} +\sqrt{\frac{N}{k}} & \text{with probability } \frac{1}{2N} \\ 0 & \text{with probability } 1 - \frac{1}{N} \\ -\sqrt{\frac{N}{k}} & \text{with probability } \frac{1}{2N} \ . \end{cases} \tag{5.46}$$

Afterwards, M serves as the linear transformation between spaces of original and reduced dimensionality, i.e.

$$\boldsymbol{y} = M\boldsymbol{x} \ . \tag{5.47}$$

Agglomerative Feature Clustering (AGG) Agglomerative Feature Clustering is a dimensionality reduction approach that merges data along the feature dimension until a target dimensionality is reached. To do so, we will consider $D \in \mathbb{R}^{n \times f}$ the data matrix of n samples of length f each. To reduce the number of samples, we will consider

$$F = D^T \tag{5.48}$$

to be the feature matrix. We can then define an agglomeration function g, that will yield a measure of width of a cluster. We furthermore define a merge function $h(f_i, f_j)$ that will merge two given vectors. We will then, iteratively, select two rows of F, f_i and f_j as

$$\underset{f_i, f_j}{argmin}\ h(f_i, f_j) \tag{5.49}$$

to be subsequently merged to $g(f_i, f_j)$. The resulting merged vector then replaces f_i and f_j in F. This procedure is repeated until a target feature dimensionality k is reached. We can then extract our transformed feature set as

$$D_k = F_k^T \ . \tag{5.50}$$

New samples will then be transformed into the agglomerated feature space by repeating the merging step performed in obtaining F_k. For the remainder of this thesis, we will use

$$g(f_i, f_j) := \sqrt{f_i^2 + f_j^2} \tag{5.51}$$

$$h(f_i, f_j) := \left\|f_i - f_j\right\|_2 \tag{5.52}$$

when agglomerative clustering is used.

KMeans (KM) K-means is a clustering algorithm that attempts to construct k clusters of equal variance within the data, i.e. picking an optimal set K^* of k means μ_j, such that

$$K^* := \underset{K}{argmin} \left(\sum_{i=1}^{n} \min_{\mu_j \in K} \|x_i - \mu_j\|^2 \right) , \tag{5.53}$$

yielding a Voronoi-segmentation of the original data space. Using that, we can define a transformation

$$T(x) = \sum_{i=1}^{k} \|x - \mu_i\| \tag{5.54}$$

that represents a data sample x with respect to its distances to the k means that have been computed in the previous step, yielding a k-dimensional representation of a higher dimensional sample.

Linear Discriminant Analysis (LDA) Linear Discriminant Analysis as a data transformation technique can be understood as finding a representation of the data that attempts to maximize inter-cluster distances while also attempting to minimize within-cluster distances.

To obtain those, we first need to identify clusters. There are two options of identifying clusters, one being supervised, clustering the data beforehand according to their label, the second one being unsupervised, by setting a number of clusters to identify, possibly retaining more dimensions if the number of clusters is chosen larger as the number of classes. However, while this might result in multiple clusters of the same label, as long as the number of clusters is larger than the number of labels, this is expected to retain clusters of labels. To attempt an unsupervised approach with a target dimensionality k, one can first employ a k-means strategy as described above to identify clusters in the data with their centers μ_i. Afterwards, one can compute the summed inter-cluster variance that shall be maximized, as

$$\sigma_{inter} = \sum_{i=1}^{k} N_i (\mu_i - \mu_D)(\mu_i - \mu_D)^T \tag{5.55}$$

with μ_D being the overall center of the dataset and N_i being the cluster size of the respective clusters, to ensure equivalent weighting.

Furthermore, we can establish the intra-cluster variances in the form of

$$\sigma_{intra} = \sum_{i=1}^{k} \sum_{x \in D_i} (x - \mu_i)(x - \mu_i)^T \tag{5.56}$$

which is the sum of the variances of the individual clusters.

We now have established measures for inter-cluster variance as well as intra-cluster variance. We now want to find a representation such that inter-cluster variances are maximized, while intra-cluster variances are minimized. To do so, we will compute the discrimination matrix

$$S = \sigma_{inter} \sigma_{intra}^{-1} \tag{5.57}$$

and solve its eigenvalue problem to obtain an eigenvector representation of the original dataset. As we wanted to maximize the property represented by S, we then can leverage the k eigenvectors with the largest corresponding eigenvalues to find a projection W that best fulfills this property, yielding a new subspace representation of the original dataset as well as additional samples that shall be transformed onto this subspace that is dimensionality reduced such that inter-cluster

distances are pronounced, while intra-cluster distances are reduced. A transformation of an arbitrary point from the space of the original dataset then simply reads

$$\boldsymbol{x_i} \mapsto W\boldsymbol{x_i} \ . \tag{5.58}$$

5.4. Classification schemes

The resulting space of samples will then be used as a basis space for a set of supervised machine learning algorithms. Different algorithms employ different strategies, and a priori it is unclear, especially for a research question that has not been attempted before, which of those is the best choice. Hence, to determine this best choice for the data that we have obtained, we will include several candidates into the investigation.

We will now briefly review the underlying procedures that are used for the classification task itself. Among these, we identify two particular groups of classifiers, the tree-based schemes that contain the Decision Tree, the Random Forest and the Extra Tree scheme, as well as the Bayes-based classifiers, who leverage Bayes' law to facilitate classification. These are the Naive Bayes, the Quadratic Discriminant Analysis and Gaussian Processes. In addition to these two groups we will investigate the suitability of of the k-Nearest-Neighbours, Passive Aggressive and AdaBoost schemes, which we will review first:

k-Nearest-Neighbours The k-Nearest-Neighbours algorithm determines the class of a sample based on the closest surrounding samples. It has already been introduced in the previous part, hence see chapter 2.1.

Passive Aggressive Online Passive-Aggressive algorithms constitute a class of linear algorithms with an aggressive weight update step as well as a wide range of applications, originally developed by [**?**]. In this thesis we will focus on the subset necessary for applying this class of algorithms to classification specifically.

To do so, we will first define the binary classification problem, i.e. a sample $x \in \mathbb{R}^n$ having a label y_t that can be one of two values, so without restriction of generality we define

$$y_t \in +1, -1 \ . \tag{5.59}$$

We further define a vector of weights $w \in \mathbb{R}^n$, such that we can predict a label

$$y_p = sign(w \cdot x) \ . \tag{5.60}$$

If $y_t y_p \geq 1$, the algorithm has made a correct prediction, with $\|w \cdot x\|$ being a measure of confidence into this prediction. Ideally, we want a correct prediction with a high confidence, i.e. $y_p(w \cdot x) \geq 1$. If that is not the case, we need to adjust the weights. To determine an adjusted weight vector, we first define a loss function to be

$$\ell(w, (x, y)) := \begin{cases} 0 & \text{if } y_t(w \cdot x) \geq 1 \\ 1 - y_t(w \cdot x) & \text{otherwise} \end{cases} \tag{5.61}$$

with (x, y) denoting a sample with attached label as well as a given weight vector w.

In the next step, we formulate an update rule to the weight vector w based on the loss defined above.

We define the initial weight vector to be

$$w = (0, \cdots, 0)^T \tag{5.62}$$

and we define the update rule for the weight vector to be

$$w_i = \underset{w \in \mathbb{R}^N}{argmin} \frac{1}{2} \|w - w_{i-1}\|^2 \text{ s.t. } \ell(w, (x_i, y_i)) = 0 \ . \tag{5.63}$$

This way, we find the weight vector w_i to be closest to the previous vector of weights w_{i-1}, i.e. the vector with minimal necessary adjustment, that doesn't yield any loss on the new sample (x_i, y_i).

If the current vector of weights already does not yield any loss for the incoming sample (x_i, y_i), the update rule (5.63) is trivially fulfilled by

$$w_i = w_{i-1} \ , \tag{5.64}$$

the algorithm hence stays *passive*, yielding the first part of the name of this class of algorithms. If this is not fulfilled, (5.63) it *aggressively* adapts the weight vector to losslessness for each new sample, yielding second part of the name of this class of algorithms.

Theorem 4. *The update rule*

$$w_i = \underset{w \in \mathbb{R}^N}{argmin} \frac{1}{2} \|w - w_{i-1}\|^2 \ \ s.t. \ \ell(w, (x_i, y_i)) = 0 \tag{5.65}$$

has a closed form solution

$$w_i = w_{i-1} + \tau_i y_i x_i \ \ with \ \tau_i = \frac{\ell_i}{\|x_i\|^2} \tag{5.66}$$

Proof.

Theorem 5. *Karush-Kuhn-Tucker conditions*
Be

$$\min_{x \in \mathbb{R}^n} F(x) \quad F : \mathbb{R}^{n_x} \to \mathbb{R} \tag{5.67}$$

$$s.t. \ G(x) = 0 \quad G : \mathbb{R}^{n_x} \to \mathbb{R}^{n_G} \tag{5.68}$$

a nonlinear optimization problem with equality constraints $G(x) = 0$. Be

$$F = x \in \mathbb{R}^n | G(x) = 0 \tag{5.69}$$

the valid set of solutions for this optimization problem. Be $x^* \in F$ a local minimum of (5.68) *Then the Karush-Kuhn-Tucker conditions state that*

$$\forall \mathcal{L}(x, \lambda) := F(x) - \lambda^T G(x) \exists \lambda^* s.t. \tag{5.70}$$

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla_x F(x^*) - \nabla_x G(x^*) \lambda^* = 0 \ . \tag{5.71}$$

[6]
Be

$$\mathcal{L}(w, \tau) = \frac{1}{2} \|w - w_{i-1}\|^2 + \tau \left(1 - y_i \left(w \cdot x\right)\right) \tag{5.72}$$

the Lagrangian of the optimization problem (5.63) and $\tau \geq 0$. We further observe, that the objective function of (5.63) is convex. Hence, we furthermore observe that the Lagrangian of this optimization problem has one feasible affine constraint, hence Slater's condition [25] guarantees equivalence of the optimum to the satisfaction of the Karush-Khun-Tucker conditions in the

general case and consequentially also for the case of absence of inequality constraints as stated in Theorem 5. This allows us to conclude that we can find an optimum via the first derivative only, hence

$$0 = \nabla_w \mathcal{L}(w, \tau) = w - w_i - \tau y_i x_i \tag{5.73}$$

$$\Rightarrow w = w_i + \tau y_i x_i \tag{5.74}$$

and leverage this finding in (5.72) to obtain

$$\mathcal{L}(\tau) = -\frac{1}{2}\tau^2 \left\|x_i\right\|^2 + \tau \left(1 - y_i \left(w \cdot x\right)\right) \ . \tag{5.75}$$

Finding the optimum of this Lagrangian in the τ-dimension yields

$$0 = \frac{\partial \mathcal{L}(\tau)}{\partial \tau} \tag{5.76}$$

$$= -\tau^2 \left\|x_i\right\|^2 + \left(1 - y_i \left(w \cdot x\right)\right) \tag{5.77}$$

$$\Rightarrow \tau = \frac{\left(1 - y_i \left(w \cdot x\right)\right)}{\left\|x_i\right\|^2} \ . \tag{5.78}$$

With definition (5.61), we find

$$\tau = \frac{\ell_i}{\left\|x_i\right\|^2} \tag{5.79}$$

Combining this with the intermediate result

$$0 = \nabla_w \mathcal{L}(w, \tau) \Rightarrow w = w_i + \tau y_i x_i \tag{5.80}$$

we find the postulated closed form solution for our update rule to be

$$w_i = w_{i-1} + \tau_i y_i x_i \text{ with } \tau_i = \frac{\ell_i}{\left\|x_i\right\|^2} \ . \tag{5.81}$$

□

The current construction of the algorithm employs, as discussed, a very aggressive update strategy for the weight vector, if classification errors occur. This, however, could lead to overaggressively weighting later samples compared to earlier ones. To reduce this, we introduce a slack variable ξ into the update rule (5.63), controlled by an aggressiveness parameter C, so that we obtain an updated update rule

$$w_i = \underset{w \in \mathbb{R}^N}{argmin} \frac{1}{2} \left\|w - w_{i-1}\right\|^2 + C\xi \text{ s.t. } \ell(w, (x_i, y_i)) = 0, \xi \geq 0 \ . \tag{5.82}$$

with an additional condition

$$\tau_i = \min \left\{ C, \frac{\ell_i}{\left\|x_i\right\|^2} \right\} \ . \tag{5.83}$$

A passive-aggressive classifier using this type of aggressiveness control is referred to as *PA-I* and will be used for the remainder of this analysis with

$$C = \xi = 1 \ . \tag{5.84}$$

Alternatively, the authors propose an alternative formulation for the update rule,

$$w_i = \underset{w \in \mathbb{R}^N}{argmin} \frac{1}{2} \|w - w_{i-1}\|^2 + C\xi^2 \text{ s.t. } \ell(w, (x_i, y_i)) = 0 \ , \tag{5.85}$$

which allows to omit the condition $\xi \geq 0$ and is carries the name *PA-II* and employs the additional condition

$$\tau_i = \frac{\ell_i}{\|x_i\|^2 + \frac{1}{2C}} \ . \tag{5.86}$$

We will, focus the analyses made in this thesis on *PA-I* and will abstain from using *PA-II*.

Until this point, we have discussed binary classification using the passive-aggressive approach. We will now generalize this approach to classifying L labels with $L \geq 2$.

To do so, we will define a set of L weight vectors w_l, each of them associated with a label y_l. Each of these combinations (w_l, y_l) constitutes an individual, independent passive-aggressive classifier as discussed above. For every combination of samples $(x_i, y_i), y_i \in (y_1, \ldots, y_L)$, we then compute the *score*

$$s_l = w_l \cdot x_i \tag{5.87}$$

for each sample x_i for the different classifiers and select the classification result y_{pred}

$$y_{pred} = y_{k^*} \text{ s.t. } k^* = \underset{k \in (0,\ldots L)}{argmax} \ s_k \ , \tag{5.88}$$

i.e. the label with the highest score. Weight updates will be performed on each individual PA-classifier individually by locally defining

$$y_{local} = +1 \text{ if } y_i = y_l \tag{5.89}$$

$$y_{local} = -1 \text{ if } y_i \neq y_l \tag{5.90}$$

Adaboost Adaboost leverages an arbitrary classification algorithm and boosts its precision by successively weighting the input samples according to their difficulty to be classified and was initially proposed in its current form by [26].

To do so, we will first assume the existence of a basic classification scheme that is used for classification of a given dataset. While Adaboost is capable of boosting any classification scheme, it is typically combined with a more simple classifier, such as a simple linear split. Such a classification scheme C, when presented with a dataset, will be trained on that dataset with equal relevance and hence weight on all samples. Consequentially, the accuracy α of said classifier C can be computed as

$$\alpha = \frac{\text{number of correctly classified samples}}{\text{number of all samples}} \tag{5.91}$$

and its error consequentially as

$$e_C = 1 - \alpha \tag{5.92}$$

This way, we can asses the effectiveness of a classifier, which in this thesis will be defined as

$$\mu = log\left(\frac{1 - e_C}{e_C}\right) \tag{5.93}$$

By computing this effectiveness, we can then define the an adjusted measure of relevance $w_{t,i}$ for each sample x_i at iteration t that was *incorrectly* classified via

$$\tilde{w}_{t+1,i} = w_{t,i} e^{\mu} \ . \tag{5.94}$$

and renormalizing all weights to the total sum of weights, i.e.

$$\boldsymbol{w}_{t+1} = \frac{\tilde{\boldsymbol{w}}_{t+1}}{\sum_{i=1}^{n} w_{t,i}} \tag{5.95}$$

By doing so, we increase the importance that is attributed to the incorrectly classified samples in iteration $t+1$, hence making it more likely that they will be classified correctly when setting up a classification scheme. Afterwards, the samples in the dataset to be classified get (re-)weighted by $\boldsymbol{w}_{t+1}$ and the initial classification scheme is reapplied to the dataset, yielding a new classifier C_{t+1} with a potentially different effectiveness mu_{t+1}.

The same procedure can then be reapplied to obtain an ensemble C_t of successively weight-optimized realizations of the leveraged classification scheme, yielding increasing sensitivity to samples that have turned out to be difficult to classify. After obtaining such an ensemble, a new sample x_* is classified by obtaining a prediction $y_t, t \in [1, \ldots, T]$ of every classifier in the ensemble, then taking the label with the highest resulting occurrence, balancing more general classifiers from the start of the boosting procedure and classifiers more specific to difficult-to-classify samples from later in the boosting process.

For practical purposes, this thesis uses a linear separator as the simple classifier that is boosted, which is implemented as a decision tree classifier (see section 5.4.2) of depth 1. The specific implementation used corresponds to Adaboost-SAMME as proposed by [38], which is designed for multiclass classification. To be suited for $k > 2$ classes, we modify the measure of effectiveness of AdaBoost, (5.93), to

$$\mu = log\left(\frac{1 - e_C}{e_C}\right) + log(k-1) \ . \tag{5.96}$$

For two classes, i.e. $k = 2$, this equation trivially simplifies to eq. (5.93), whereas for $k > 2$ a classifier's effectiveness measure is increased if more classes are present. This is on the one hand intuitively plausible as a classifier that achieves $\frac{1}{N} + \epsilon$ when N classes are available will already be an improvement over random guessing, whereas it becomes increasingly more difficult to achieve this the more classes and hence options for misclassification exist, on the other hand it also follows as a natural result of considering a forward stage-wise additive model with a multi-class exponential loss function. For a detailed derivation of eq. (5.96) refer to [38].

5.4.1. Bayes-based classifiers

Naive Bayes The Naive Bayes classification scheme is based on Bayes' theorem, which in its simplest form states the probability of an event y assumed an event or condition x as

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} \ , \tag{5.97}$$

which generalizes straightforwardly to

$$p(y|x_1, \cdots, x_n) = \frac{p(y)p(x_1, \cdots, x_n|y)}{p(x_1, \cdots, x_n)} \tag{5.98}$$

if more than one condition x is at play for the probability of y. We can now identify x_1 to x_n as n different features and consequentially our classification problem is the determination of a label y given a feature vector $(x_1, \cdots, x_n)$. To compute our conditional probability for a label y we need to compute the three components $p(y)$, $p(\boldsymbol{x}|y)$ and $p(\boldsymbol{x})$. Firstly, we will investigate the probability of a sample given a specific label. We will assume independence of components, i.e.

$$p(x_1, \cdots, x_n|y) = \prod_{i=1}^{n} p(x_i|y) \ , \tag{5.99}$$

as [103] among others have stated that and investigated why a Naive Bayes classification procedure exhibits robust classification performance under this assumption even if it is violated, therefore we will still consider this method for application even if the underlying data's features have a high likelihood of not being independent. $p(x_i|y)$ can then be determined by a distribution assumption. Assuming all data samples are mostly centered, and the outward flanks are rapidly decreasing, we can employ a Laplace approximation.

Definition 3. *The **Laplace Approximation** of a probability distribution $p(x)$ assumes that this distribution's cumulants of order higher than 2 are negligible in size.*

Theorem 6. *The Laplace approximation of a probability distribution $p(x)$ is equivalent to approximating said distribution with a normal distribution, i.e.*

$$p(x) \approx \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \quad . \tag{5.100}$$

Theorem 7. *For every probability distribution $p(x)$, there exists a characteristic function*

$$\Phi(t) = \int p(x)e^{-itx} \tag{5.101}$$

that completely defines the corresponding probability distribution.

Proof. Theorem 7
Be $p(x)$ a probability function and

$$\Phi(t) = \int p(x)e^{-itx}dt \tag{5.102}$$

its corresponding characteristic function. Then $p(x)$ is completely defined by $\Phi(t)$ as it can be reconstructed without loss by reapplying a Fourier-transformation, i.e.

$$\Phi(t) = \int p(x)e^{-itx}\mathrm{d}x \tag{5.103}$$

$$\Leftrightarrow p(x) = \int \frac{1}{2\pi}\Phi(t)e^{itx}\mathrm{d}t \tag{5.104}$$

and consequently completely defines the underlying probability distribution. □

Proof. Theorem 6

Be $Phi(t)$ the characteristic function of the probability distribution $p(x)$. The cumulants κ_n that are obtained by Taylor-expanding the logarithm of $\Phi(t)$, i.e.

$$ln\left(\Phi\left(t\right)\right) = \sum_{n=1}^{\inf} \frac{(it)^n}{n!}\kappa_n \tag{5.105}$$

define the probability distribution $p(x)$ completely (see theorem 7) as they define the characteristic function completely, as

$$\Phi(t) = exp\left(\sum_{n=1}^{\inf} \frac{(it)^n}{n!}\kappa_n\right) \tag{5.106}$$

holds. Furthermore, we know that for Gaussian distributions

$$\kappa_n = 0 \; \forall \; n \geq 3 \tag{5.107}$$

equally holds. Hence, assuming $\kappa_n \approx 0$ for $n > 2$ for a distribution $p(x)$, this distribution is equivalent to a Gaussian distribution, and can consequentially be described by its probability density function

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \; . \tag{5.108}$$

□

Applying a Laplace approximation, we hence approximate the distribution $p(x_i|y)$ as

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(\frac{-(x_i-\mu_y)^2}{2\sigma_y^2}\right) \tag{5.109}$$

By obtaining μ_y and σ_y via fitting of this distribution to the training data available that is labeled with the according label y, we can compute the multivariate probability distribution of $\boldsymbol{x}$ with respect to the existing training data for all labels present. Furthermore, we need to compute the probability of a label on its own, $p(y)$. Without further indication, we will approximate the probability of $p(y)$ by its prevalence in the training set. The last property we need to compute is $p(\boldsymbol{x})$. Given that we want to apply a probability estimate for any arbitrary point, there is no probability assumption on $p(\boldsymbol{x})$ other than that every point is equally probable, making it a constant for our procedure.

After we have now established a measure of probability for a label, and hence a measure of confidence, the decision for a label $y \in (0, 1)$ is then established by defining our classifier to assign label 1 if

$$c_1(\boldsymbol{x}) = \frac{p(1|\boldsymbol{x})}{p(0|\boldsymbol{x})} = \frac{p(y=1)}{p(y=0)} \prod_{i=1}^{n} \frac{p(x_i|y=1)}{p(x_i|y=0)} \geq 1 \tag{5.110}$$

and equally to assign label 0 if

$$c_0(\boldsymbol{x}) = c_1(\boldsymbol{x})^{-1} > 1 \; . \tag{5.111}$$

To modify eq. (5.112) to accommodate for more than two classes, one can either opt to simply assign the label with the highest confidence, regardless of its value, to obtain a classifier that will always attempt to classify a sample, or to add an additional constraint η to the confidence, either as a constraint onto the confidence itself or on a classification confidence estimate, such as

$$c_{y_i}(\boldsymbol{x}) = \frac{p(y_i|\boldsymbol{x})}{\sum_{j=1, j\neq i}^{J} p(y_j|\boldsymbol{x})} \geq \eta \geq \frac{1}{J} \; . \tag{5.112}$$

Quadratic Discriminant Analysis Quadratic Discriminant Analysis is an modification of the Naive Bayes approach presented in the last section, but employs a different approximation of the probability measure $p(\boldsymbol{x}|y)$. Instead of assuming dimensionality independence, Quadratic Discriminant Analysis will instead assume a full covariance matrix for each class, while still being fundamentally Gaussian distributed. The probability density function then reads as

$$p(\boldsymbol{x}|y) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma_y|^2}} exp\left(-\frac{(x_i-\mu_y)^T \Sigma_y^{-1} (x_i-\mu_y)}{2}\right) \tag{5.113}$$

with $\boldsymbol{x} \in \mathbb{R}^d$ and Σ_y denoting the covariance matrix of class y. In case

$$\Sigma_y = \Sigma \; \forall \; y \; , \tag{5.114}$$

i.e. all classes have the same covariance matrix, Quadratic Discriminant analysis simplifies to Linear Discriminant Analysis. If the covariance matrices Σ_y are all diagonal, i.e. the components of the feature vector $\boldsymbol{x}$ are independent, we can deduce

$$p(\boldsymbol{x}|y) = \prod_{i=1}^{n} p(x_i|y) \; , \tag{5.115}$$

which is the reemergent base assumption of a Naive Bayes classification approach, and consequentially QDA is simplifies to that.

Gaussian Processes A Gaussian Process attempt to predict the corresponding value y_* of an additional sample x_* by assuming a normal distribution of all samples $X = (x_1, \cdots, x_n)$, and consequentially predict y_* to be the most likely occurrence given the known data. The given data X can be described by their values $(y_1, \cdots, y_n)$, but also by the covariance matrix of the distribution

$$K_{i,j} = cov(x_i, x_j) \tag{5.116}$$

with $K_{i,i}$ describing the variance or uncertainty of the sample x_i. The fundamental second assumption that forms the base of a Gaussian Process is the existence of a *covariance function* $k(x_1, x_2)$, that is capable of describing said covariance, such that

$$K_{i,j} = k(x_i, x_j) = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \cdots & \cdots & \cdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \; . \tag{5.117}$$

This covariance function k can be infused with the fundamental modelling assumptions of the model, if existing, such as periodicity if the problem is periodic, such as

$$k(x_1, x_2) \approx k(x_1, x_2 + \pi) \; . \tag{5.118}$$

If nothing specific is known about the underlying data, a classical approach can be the base assumption, that samples further apart should be less correlated than samples close by, i.e.

$$\|x_1, x_2\| > \|x_1, x_3\| \Rightarrow k(x_1, x_2) > k(x_1, x_3) \; . \tag{5.119}$$

Specific options for such a k include a Gaussian function kernel or a radial basis function kernel, but in principle, any bell curve would principally be suited for this purpose. This thesis specifically will leverage a radial basis function of kind

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i, x_j\|_2^2}{2\theta^2}\right) \tag{5.120}$$

for the following investigations. The parameter θ indicates a typical length of the problem, i.e. on which length scale correlations of points are still high and where correlation starts to fade away. It can be determined based on the input data during the training procedure. The notations $k(x_i, x_j)$ and $k(x_i, x_j, \theta)$ will be used interchangeably, depending on the relevance of the parametrization θ to the point being discussed.

We will now introduce a new sample point x_*, whose value y_* is to be predicted. By defining

$$K_* = [k(x_*, x_1), \cdots, k(x_*, x_n)] \tag{5.121}$$
$$K_{**} = [k(x_*, x_*)] \tag{5.122}$$

(5.117) can be extended to

$$\tilde{K} = \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix} . \tag{5.123}$$

We can then conclude that by construction of this method our sample including our added point (x_*, y_*) is drawn from a multivariate (by assumption zero centered) Gaussian distribution, i.e.

$$\begin{bmatrix} \vec{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) . \tag{5.124}$$

We are now interested in the value y_* for the new sample based on the information of the given data, which itself is Gaussian distributed [21]:

$$y_* | \vec{y} \sim \mathcal{N}(K_* K^{-1} \vec{y}, K_{**} - K_* K^{-1} K_*^T) , \tag{5.125}$$

yielding a most likely estimate

$$y_* = K_* K^{-1} \vec{y} \tag{5.126}$$

with a variance

$$var(y_*) = K_{**} - K_* K^{-1} K_*^T \tag{5.127}$$

that can be used as a confidence estimate for the estimate. An example of such an application of a Gaussian Process is depicted exemplarily in Fig. 5.3.

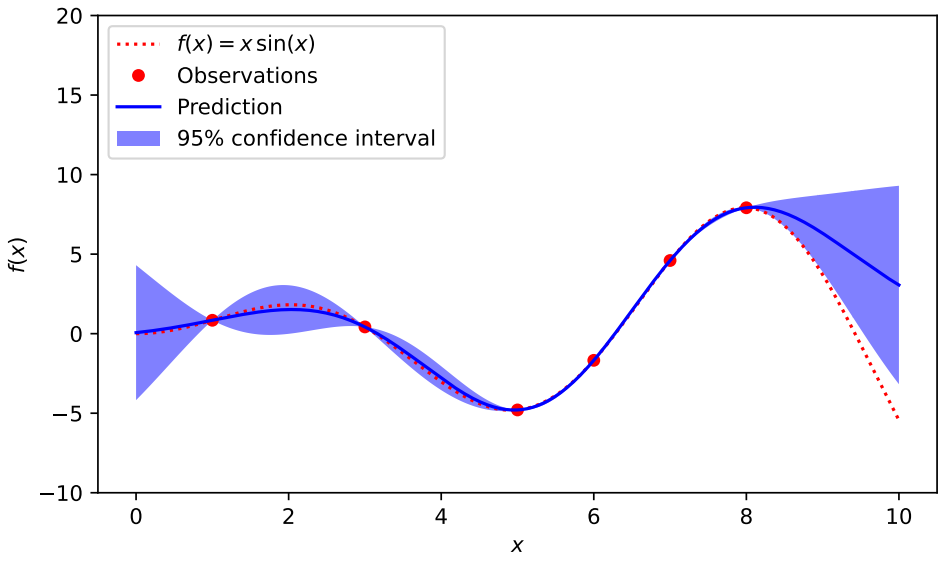

Figure 5.3.: Exemplary Gaussian Process regression for $f(x) = x\ sin(x)$ based on [20].

So far, this procedure allows to compute regressions with Gaussian Processes. To extend this to classification, we will first define two classes with values 0 and 1. A classification procedure

is then in first approximation conceptually a Gaussian Process regression with $y_i \in (0,1) \forall i$. However, two more modifications will be made: We will introduce a probability measure

$$\pi : \mathbb{R} \to [0,1] \tag{5.128}$$

$$x \mapsto \frac{1}{1+e^x} \in [0,1] \ , \tag{5.129}$$

which is depicted in Fig. 5.4 This will allow us to map the result of the previously discussed

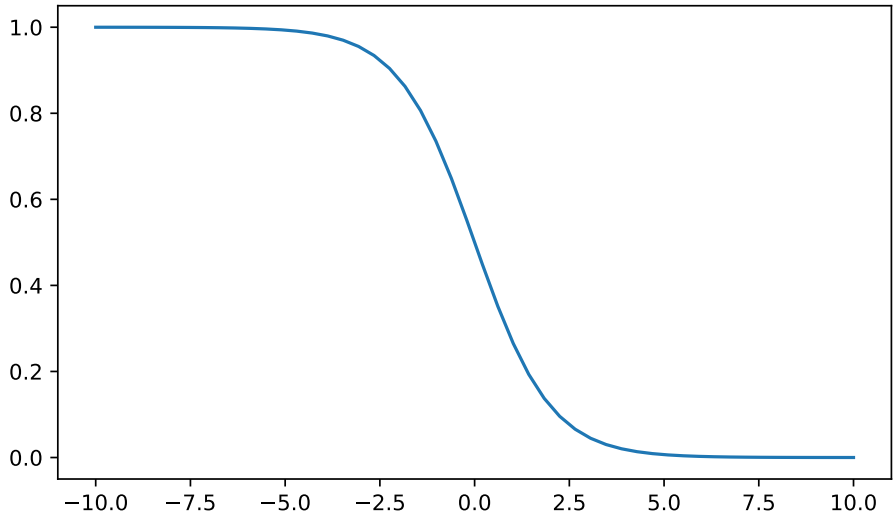

Figure 5.4.: Depiction of the standard logistic function serving as $\pi(y_*)$. x-axes depicts arbitrarily scaled numbers, y-axis will be used for probabilistic weighting.

regression to the probability of our result having one of the two specified labels, without loss of generality p is constructed such that $p(y_*)$ describes the probability of the newly introduced sample x_* being of class 1. To find the most likely probability for x_* being of class 1, we can compute the expectation value

$$\pi_* = \int p(y_*)(y_*|\vec{y})dy_* \ , \tag{5.130}$$

which, according to [98], is solved by

$$\pi_* = \Phi\left(\frac{\bar{y_*}}{\sqrt{1+var(y_*)}}\right) \ , \tag{5.131}$$

yielding our best probability estimate for x_* being of class 1. However, as we are marginalizing over all possible y_* in eq. (5.130), we need to consider that this has implications onto the variance of y_*. In fact, we want to maximize the probability of

$$p(\vec{y_*}|\vec{x},\vec{y}) = \frac{p(\vec{y}|\vec{y_*})p(\vec{y_*}|\vec{x})}{p(\vec{y}|\vec{x})} \ . \tag{5.132}$$

Assuming integral independence and equal distribution of the dataset, i.e.

$$p(\vec{y}|\vec{f}) = \Pi_{i=1}^{n} p(y_i|f_i) \ , \tag{5.133}$$

then by construction we find $p(y|f) = \pi(f)$. We furthermore want to find $p(\vec{f}|\vec{x})$ such that eq. (5.132) gets maximized. This is the case when the derivative of eq. (5.132) with respect to $\vec{f}$ is zero, and when it's equivalent logarithm's derivative is zero, which leads [21] to find

$$y'_* = K\nabla log(p(\vec{y}|\vec{y'_*}) \ , \tag{5.134}$$

which can be solved iteratively. Furthermore, we need to compute the corresponding variance, which we find by taking the logarithm of eq. (5.132) and computing the negative of its second derivative [21]. This yields

$$var(\vec{y_*}) = (K^-1 - \Delta log(p(\vec{y}|\vec{y_*})))^{-1} \ . \tag{5.135}$$

The resulting distribution is no longer Gaussian, however, we will apply a Laplace approximation, i.e. assume it still is [98], hence

$$p(\vec{y_*}|\vec{x}, \vec{y}) \sim \mathcal{N}(\vec{y'_*}, (K^-1 - \Delta log(p(\vec{y}|\vec{y_*})))^{-1} = \mathcal{N}(\vec{y'_*}, K') \ . \tag{5.136}$$

This leads to a modified, but more accurate version of eq. (5.125) in the form of

$$y_*|\vec{y} \sim \mathcal{N}(K_* K^{-1} \vec{y'_*}, K_{**} - K_*(K')^{-1} K_*^T) \ , \tag{5.137}$$

considering the effect of marginalization.

To extend Gaussian Process classification to more than two classes, we proceed similar to the Passive-Aggressive classification algorithm previously, by fitting one Gaussian Process per class, the results of which get concatenated and paired with a modified y, which is 1 for each Gaussian Process that was fitted to its class and 0 otherwise. Based on this new, dimensionality-extended data set, a Gaussian Process classifier can be derived similar to what has been described above. For details about the procedure, refer to [98].

5.4.2. Tree-based Methods

Decision Tree While there is a variety of different implementations of decision trees, such as the Iterative Dichotomizer 3 [77], its successor Iterative Dichotomizer 4.5 [78] or its newest iteration 5.0 [93], this thesis will leverage the Classification and Regression Trees (CART) model type, originally introduced by Breimann et al. [9].

To classify a sample s using a such a Decision Tree classifier, we will first construct A CART-type decision tree model consists of the construction of a binary tree B.

Be $D = (d_1, \cdots, d_n)$ a set of samples to be classified with sample labels $(l_{d_1}, \cdots, l_{d_n})$, and be $\eta = (L_1, \cdots, L_m)$ the set of labels, such that

$$l_{d_i} \in \eta \ \forall \ d_i \in D \ . \tag{5.138}$$

To classify a new sample s with a decision tree classifier, we will first split dataset D into two datasets D_1 and D_2 by a splitting criterion c. This splitting criterion $c \in C$, with C being the set of possible splits, is designed to, without loss of generality, split D, such that purity of D_1, $p_1(D, c)$ is maximized:

$$c = \underset{c \in C}{argmin} \ p_1(D, c) \tag{5.139}$$

This process can then afterwards be repeated for the datasets D_1 and D_2, which themselves can be split by the same criterion, to obtain datasets $D_{1,1}$ and $D_{1,2}$ as descendants of dataset D_1 as well as $D_{2,1}$ and $D_{2,2}$ obtained from splitting D_2. The process can be repeated either until a datasubset reaches absolute purity, i.e. containing only samples of a single class, or alternatively until a maximum number of dataset splits is reached. A maximum constraint on the number of splits is typically employed to reduce the impact of overfitting.

When discussing purity of a dataset, we first need to establish a measure of purity. One of the commonly used measures of to do so in literature are Gini-impurity [57] $G(D)$, which reads as

$$G(D) = \sum_{l_j \in \eta} p_{l_j} e_{l_j} \tag{5.140}$$

with p_{l_j} corresponding to the probability of a sample in the dataset D having label l_j and e_{l_j} to the probability of a sample not having said label l_j. We can state that

$$e_j = \sum_{l_i \in \eta l_j} p_{l_i} = 1 - p_{l_j} \tag{5.141}$$

and hence

$$G(D) = \sum_{l_i \in \eta} p_{l_i} \left(1 - p_{l_i}\right) = 1 - \sum_{l_i \in \eta} p_{l_i} \ . \tag{5.142}$$

If the probabilities inside a dataset D cannot be known precisely, they will be approximated by occurrence, i.e.

$$p_{l_j} \approx \frac{\sum_{d_i \in D | l_{d_i} = l_j} 1}{\sum_{d_i \in D} 1} \ . \tag{5.143}$$

An alternative way of quantifying impurity is computing a dataset's information theoretical entropy as introduced by [89]. In this formulation, the entropy of a dataset is given by its information content or surprisal S of a certain event E, such as the label of a randomly drawn sample, which in turn is characterized by Shannon as

$$S = log\left(\frac{1}{p_E}\right) = -log(p_E) \ , \tag{5.144}$$

such that events that are considered rarer yield a higher surprisal, or, in other words, events that are considered more unlikely, i.e. have a lower probability associated with them, have a higher information content.

The entropy H of a dataset D is then defined as the average expected level of surprisal a dataset is assumed to yield, i.e.

$$H(D) = -\sum_{l \in \eta} p_l \ log\left(p_l\right) \ . \tag{5.145}$$

Again, if probabilities within a dataset are unknown, they are approximated by their occurrence, as denoted in eq. (5.143).

While the splitting criterion can in theory be an arbitrary hyperplane, or even an even more arbitrary hypercurve, we additionally limit the space of possible splitting criteria to contain only criteria that split linearly in one dimension.

On a random circularly generated dataset, this yields classification areas that are depicted in Fig. 5.5 for a decision tree employing the Gini impurity criterion (5.142). The corresponding resulting hierarchy of splitting criteria can be seen in Fig. 5.6.

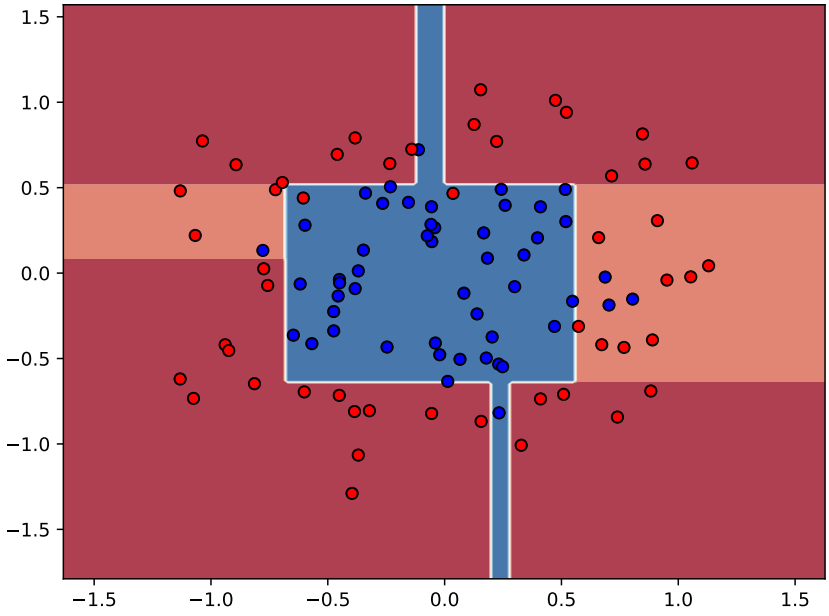

Figure 5.5.: Exemplary space split and classification regions for a two-dimensional circular randomly generated arbitrary dataset for a single CART-type Decision Tree.

After having performed the aforementioned procedure and deriving the set of splitting criteria, classification can be performed for a sample s by following the same exact splitting criteria. If those criteria, applied to s, lead to leaf in the decision tree with absolute purity, the label present in this leaf is assigned as the predicted label to s, in case of an impure leaf, the most prevalent label is assigned as a prediction. Furthermore, a measure of confidence for such a prediction can be derived from the purity of the resulting leaf according to eq. (5.143).

Random Forest A *Random Forest* is an ensemble of decision trees, firstly introduced by [43]. Its purpose is the mitigation a decision tree classifier's tendency to overfit. To achieve this, a Random Forest consists of multiple non-identical decision trees, that are derived from the same dataset. This can be achieved in different ways: Either, the dataset is split in several overlapping subsets, each of which serve as a basis to train a decision tree classifier. This way, the impact of specific samples in the dataset is reduced, as these samples should now only contribute to a subset of available trees. An alternative way of differentiating the different trees is to further constricting the set of splitting criteria C to certain dimensions. Given the comparatively small amount of base input data that we have, we will use the second differentiation variant, randomly constraining the sets C for the different trees to $M = \sqrt{N}$ of N available dimensions.

To then classify a sample s, this sample will be classified with each individual tree of the forest. The resulting label for s can then either be determined by a majority vote of the different trees, as proposed by [43], or alternatively the average confidence over all trees for the set of possible labels in question can be computed, as implemented in [8], to reduce the imbalance between trees strongly in favor of a specific label and a majority of trees just in slight favor of a different one. For the application later on, this thesis will opt for the latter.

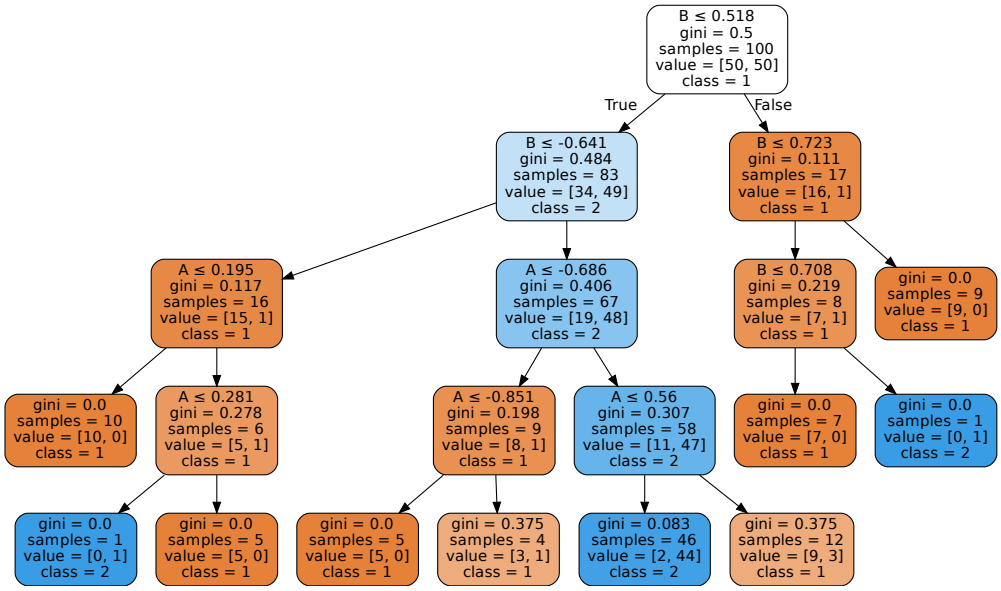

Figure 5.6.: Full decision tree used to generate decision regions in Fig. 5.5.

Extremely Randomized Trees The *Extremely Randomized Trees* are a Random Forest style classifier, which does not optimize the splitting criterion for maximum purity, but instead picks a random split for each available dimension, reducing the size of the set of possible splits C to the number of dimensions available, of which a split is then picked via the previously outlined criterion of impurity.

5.4.3. Prioring

As patients have reported during interviews to stay up to several hours in a dysfunctional cognitive state once they have entered it, we additionally leverage this information in an optional stabilization mechanism in form of adding prior information to the classification. Under the assumption that the a time window is likely to represent the same state as the previous window, we can introduce a stabilization mechanism, where the probability p_i of each sample s_i is influenced by the probabilities of the previous class probability p_{i-1} as well as the unpriored class probability $\tilde{p}_i$ in the form

$$p_i = 0.3\tilde{p}_i + 0.7p_{i-1} \ . \tag{5.146}$$

This way, previous information is included and a label change in the classification result is strongly suppressed until enough evidence is gathered to support it. With a weighting factor of 0.3 for new information, the influence of old samples with a distance a to the currently classified sample will have an influence of $(1 - 0.3)^a$. We keep the weighting factor for new information constant, but we allow the application of this methodology to be switched on or off. This, combined with a strong emphasis on the influence of the past couple of samples, will allow an assessment if such a prior stabilizes the classification or if it doesn't provide additional stability.

5.5. Assessing an optimal classification procedure

After considering all possible combinations of data characterization, feature space transformation, classification scheme and its hyperparameters, we can quickly conclude that an assessment of every possible combination becomes prohibitively costly. To further reduce the levels of our optimization problem, we will merge the given optimization dimensions into three levels in total: The **characterization level**, which consists of the selection of data fields to use, its representations and the transformation to be applied, the **classifier configuration**, which consists of the space of possible hyperparameters for a given classification algorithm and finally the training of the configured classification algorithm itself. While the optimization strategies for the last level are typically rather straightforward, we are still faced with the fact that computing all possible combinations in the remaining two levels is infeasible, therefore we will employ a Bayesian optimization strategy for them. Such a Bayesian optimization strategy is fundamentally based on Bayes law, which has been discussed beforehand already in the context of Bayesian classification algorithms. We remember that Bayes law in its most general form reads

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{5.147}$$

with two different properties A and B. In our specific case, the two properties we are looking after for the first optimization level are the accuracy A of a given classification procedure given the characterization C of said procedure, hence Bayes law for this case reads

$$p(C|A) = \frac{p(A|C)p(C)}{p(A)} \ , \tag{5.148}$$

which we want to maximize. As we are only interested in the configuration that yields the maximum accuracy, we can drop the normalization factor, hence yielding a posterior distribution for a configuration given a desired accuracy as

$$p(C|A) \propto p(A|C)p(C) \ . \tag{5.149}$$

While this is, without the normalization factor, no longer a probability distribution, it still shares the location of the optima with the original probability distribution, hence yielding all information we need to proceed. We can then sample configurations and compute their accuracy $p(A|C)$, using this to update the information we already have for the likelihood of a certain configuration C yielding a certain accuracy A to locate the optimal configuration and, more importantly, use this information as a regression point to estimate the accuracy and its level of uncertainty in the surrounding space. By doing so, we then sample additional points in that surrounding space, where the current information we already have indicates a likelihood for configurations with high accuracy. While we initially sample the space with 90 points to get first information of the distribution of accuracies given our configurations, we then proceed to leverage that information to decide where to newly acquire points given their probability of accuracy marginalized uncertainty above lower confidence bound, that is, given an expectation value μ_{C^*} and a standard deviation σ_{C^*} for the current optimal configuration, we will attempt to explore the area, where the likelihood of finding an accuracy above the lower confidence bound $\mu_{C^*} - \kappa\sigma_{C^*}$ is highest, i.e. the next configuration of interest C' to acquire new information is

$$\underset{C}{argmax} \ \mu_C - \kappa\sigma_C \ . \tag{5.150}$$

To estimate the variance and average of an arbitrary configuration we use a Random Forest regressor due to the fact that our search space does contain categorical variables, such as the

choice of transformation, which cannot be easily represented by a continuous estimator such as a Gaussian Process. With this procedure, we can now determine locations to sample from, repeatedly increasing our density distribution, to acquire enough information about our search space to make a well-informed estimate on the best configuration. The same procedure will be applied to, for a given configuration, optimize the configuration of hyperparameters for the underlying classification algorithm that is tasked with classifying the resulting dataset. The parameter κ hereby controls the amount of exploration compared to the amount of exploitation that is applied. A low κ will tend to focus on known maxima in the accuracy, preferring configurations that are close to the known best configurations, while a high κ will favor exploration and hence attempt to find optimal configurations further from the current known optimum. As we can deploy sufficient computational resources to the evaluation of the underlying objective function that is to be optimized, a single evaluation of this objective function is comparatively cheap and both the data as well as the search space is unknown and expected to be rather complex, we will choose

$$\kappa = 1.96 \tag{5.151}$$

to favor exploration over exploitation practically.

6. Investigating the impact of classification components on patient classification accuracy

6.1. Participant-specific investigations

We will now investigate the impact of the different components of this procedure on the accuracy of the classification. To ensure a sufficiently solid data basis, we will limit the investigation to participants that have participated in the study long enough such that it was possible to obtain at least eight hours of measurements that could be labeled either as a functional or as a dysfunctional state based on the patient's self-assessments. Four patients were found fulfilling this criterion and will serve as the basis for the following analysis. The data of the patients was randomly split into a training and a test set, with approximately 30% of the data serving as a test set to verify the accuracy of the resulting optimization procedure. Continuous blocks of self-assessed state were not split, the split was made only between blocks of continuous state, not inside of them. As a consequence, the split of the data into 70% and 30% is just approximate but not exact. The exact times can be found in table 6.1.

Furthermore, while the choice of classification algorithm can be part of the optimization space, we will perform separated optimizations with each classification algorithm specifically, to be able to compare their performance on the given datasets. The other characteristics will be left to be part of the upper level optimization routine, to optimize for the best configuration for a specific classifier. Furthermore, as we do have a rather complex error surface in the data and are still working with comparatively little data for what we try to achieve, we see convergence into different minima with each run. To reduce the impact of this and to still be able to evaluate the performance of different configurations, we will run the same optimization multiple times. In addition to collecting the resulting optima, we will also consider those configurations that achieve close to the maximum accuracy, because in the end we want to identify the best configuration that will perform well on multiple different patients.

patient ID	training set	test set
pat01	563 min	199 min
pat02	479 min	149 min
pat04	613 min	403 min
pat13	590 min	272 min

Table 6.1.: Amount of data used for training as well as verification of the different configurations for the single patient investigation.

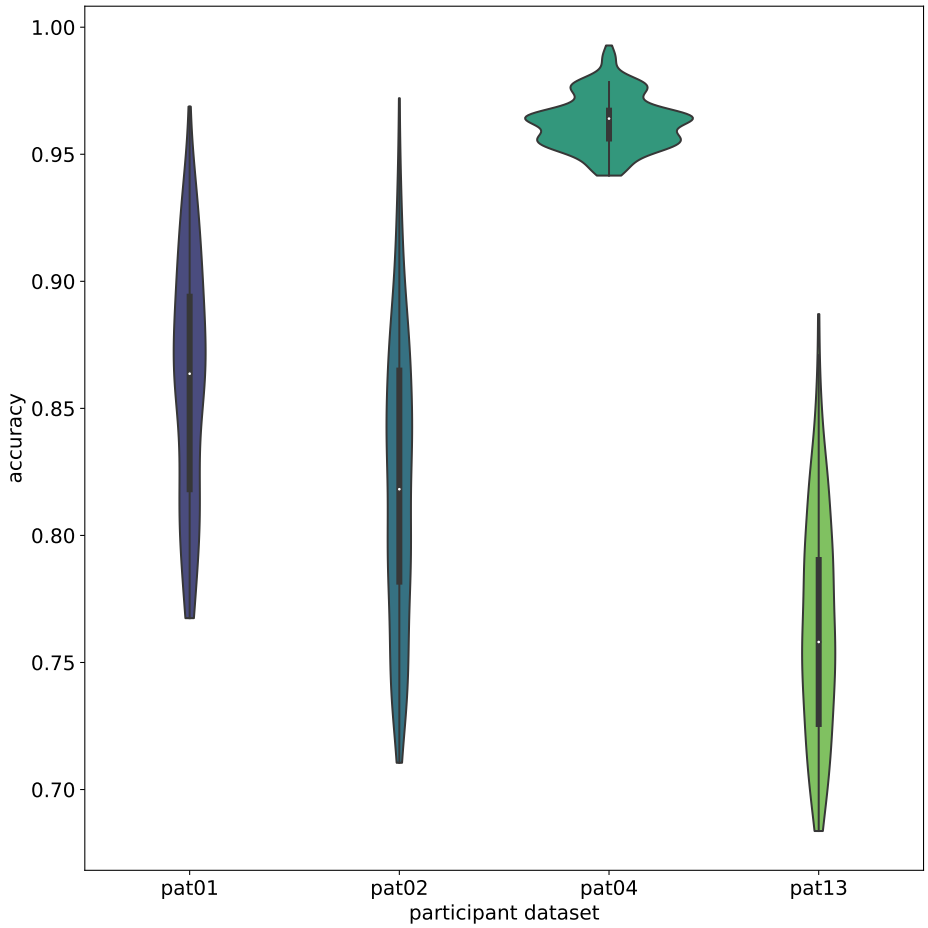

Figure 6.1.: Accuracy distribution of selected classification configurations split by participant.

6.1.1. Overall accuracy per participant

The first immediate question that arises is whether the aforementioned procedure is capable of yielding above-random accuracy at all. To investigate that question, we will sum over all classifiers and all their configurations to see the resulting accuracy distribution over the different participants. The resulting distribution can be seen in Fig. 6.1, which contains every run's optimum α^*as well as all configurations with an accuracy α that manage to be within one percentage point of α^*, i.e.

$$\alpha^* - 0.01 < \alpha < \alpha^* \ . \tag{6.1}$$

We notice that we are for all participants substantially above the 50% mark given two states that can be guessed. We furthermore see, that especially *pat04* has a untypically high accuracy. The reason for that is the non-equal distribution of labels of the different participants. So to more reasonably assess the question of improvement compared to random guessing based on the initial dataset, we need to take a look not only at the number of labels, but also their prevalence in the dataset. Based on that we can compute the accuracy that would be obtained by drawing a the classification result based on the weighted prevalence of the labels in the dataset.

patient ID	guess accuracy: training set	guess accuracy: test set
pat01	0.505	0.557
pat02	0.592	0.502
pat04	0.867	0.889
pat13	0.615	0.502

Table 6.2.: Accuracy yielded by a classification scheme that would randomly guess the label of a sample based purely on the distribution of prevalence of labels in the underlying dataset.

The corresponding numbers can be seen in Tab. 6.1. As Fig. 6.1 only shows classification accuracy based on the test set, this is what we need to compare against and we can conclude that all classification schemes gain a comfortable lead compared to the purely guesswork based classification, some configurations bridging the gap between perfect classification and random classification quite far: The median classification accuracy that could be obtained by the aforementioned selection criterion 6.1 yields an average accuracy of 0.76 with a standard deviation of 0.04 for pat13, which is an improvement of about 0.25 compared to random guesswork, bridging half the gap between absolute accuracy and weighted random classification accuracy. For pat02, the increase of accuracy is slightly bigger, with a average accuracy is about 0.82 and a standard deviation of 0.06, which bridges about 58% of that gap, for pat01 as well as pat04, $\frac{2}{3}$ and $\frac{3}{4}$ of that gap is bridged respectively at average accuracies of 0.86 and 0.96 with standard deviations of 0.05 and 0.01. Therefore we have a clear indication that the proposed optimization scheme is very capable of yielding improvements of classification accuracy over random guesswork.

6.1.2. Assessing the viability of different classification schemes

Until this point we have not differentiated yet, which configurations are actually preferable. Therefore, we will now subdivide the investigation by classification scheme, where we will merge the accuracy results of different participants, while still performing classifications separately for each participant. This allows to investigate which classification schemes perform well with the kind of data that we have at our disposal. The fact that we do not allow the classification scheme to be part of the optimization space, but instead run different optimizations for different schemes, reinforces the information we gain out of this investigation. Using the same selection criterion as in Fig. 6.1, we find that there is a slight, but noticeable difference between different algorithms, that are depicted in Fig. 6.2. The first observation is the noticeable dent that most of the distributions exhibit, in different levels of pronunciation. This is, with very high likelihood, a result of the different base accuracies we obtain for the different participants as they have already been visible in Fig. 6.1 and were discussed in the previous section. Besides that finding, we clearly see that for one, the tree based algorithms, Decision Tree, Random Forest and Extra Trees, consistently yield substantial classification accuracies, the median accuracy being above 0.87, a quarter of all accuracies reaching as high as 0.92. Given that the underlying dataset is challenging, both in terms of the implicit more than explicit patterns that we intend to detect, as well as in terms of the amount of data, which is very high given the difficulty of acquisition, but still limited in absolute terms, tree-based algorithms are expected to perform decently as they are known for reaching high accuracies on difficult datasets, albeit with an increased risk of overfitting. Interestingly, the situation is quite different for the Bayes-based classifiers: Whereas the Naive Bayes and the more complicated Gaussian Process based schemes deliver quite impressive classification performance, the Quadratic Discriminant Analysis, which attempts to employ

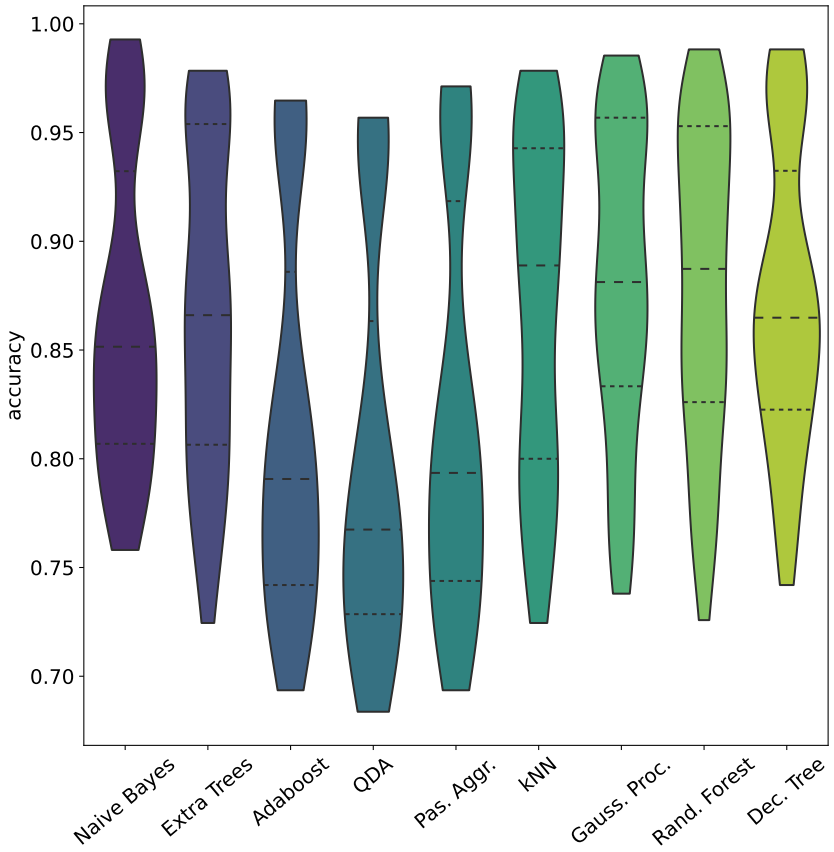

Figure 6.2.: Accuracy distribution of selected classification configurations split by algorithm. Participant-individual classification has been performed before unionization and splitting by algorithm.

a more structural probability density function, including covariance information, performs worst of all algorithms shown in virtually all criteria, the cumulative accuracy distribution is lower than any other at any point, which is quite an interesting finding.

This could indicate, that characteristics of the states that we attempt to detect are not only not well characterized by correlation, but that the set of explicit patterns that are indicative for either state contains more than one pattern per state that need to be differentiated, as correlation information is apparently not necessarily helpful, but more of a hindrance to accuracy.

We also see that the capability of the fairly simple Naive Bayes algorithm, as remarked by [103], amongst others, shows for our data as well, although the median accuracy achieved is more heavily concentrated around the accuracy area around 0.80 to 0.85, whereas the Gaussian Process is able to reach more consistently above the 0.90 mark. We further find, that the k-Nearest Neighbours algorithm also performs consistently well, with its median accuracy almost reaching 0.9 and its 25th percentile at about 0.83 and its 75th percentile at solid 0.95. The AdaBoost as well as the Passive Agressive algorithms, however, settle lower, reaching a median classification accuracy below the 25th percentile of all other classification schemes except QDA, which performs even

worse. This indicates two things: Firstly, the sub-par performance of AdaBoost could be an indicator that the complexity of the underlying dataset we are working on is somewhat evenly distributed, as boosting specific parts of the dataset doesn't seem to be a strategy that yields competitive classification accuracies. Furthermore these results indicate that attempting a linear separation might be an insufficient way of compartmentalizing the underlying data. This is also supported by the low performance of the QDA, which also indicates the more complex structure of the sample space that is not well described by its correlations.

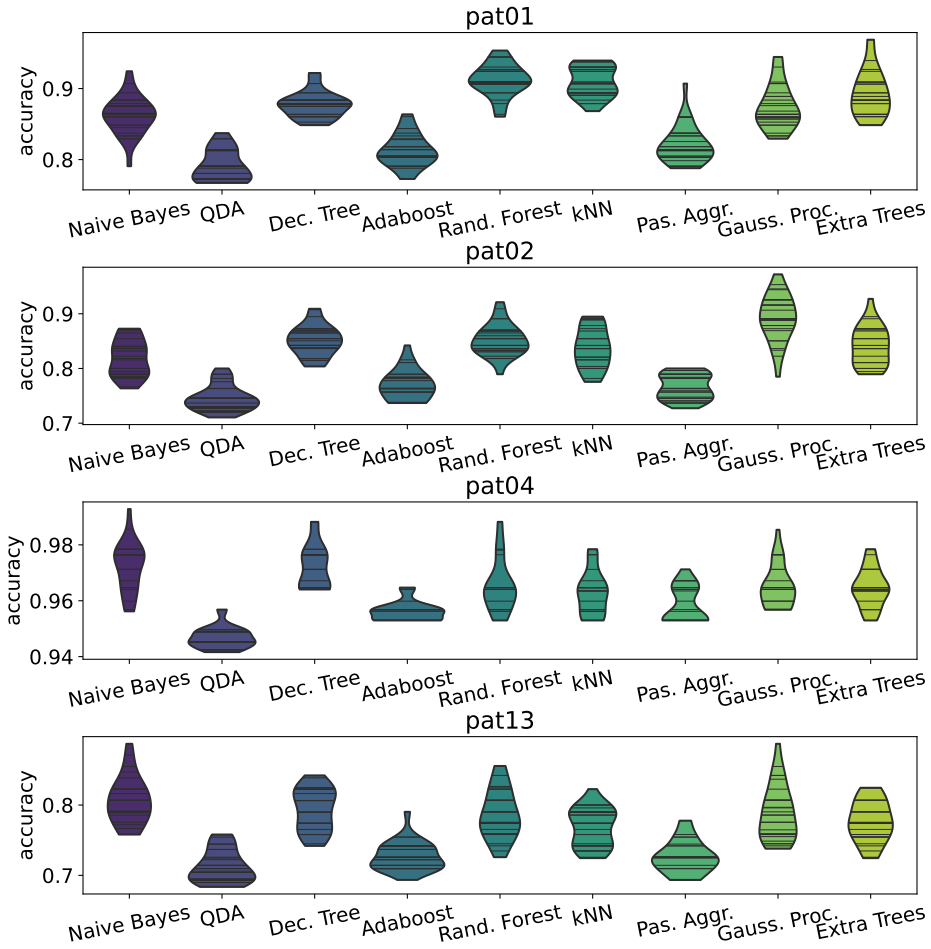

Figure 6.3.: Accuracy distribution of selected classification configurations split by algorithm and participant.

However, it still could be the case that the visible performance deficiencies, or alternatively the superior performance of some algorithms compared to other algorithms, is based on the depiction that merged accuracies obtained by different datasets, and that the top-most accuracies of, for example, the QDA, are more consistent over different participants that in other algorithms. Hence, we need to verify our findings by investigating the performance of different classification algorithms specific to different participants, which is depicted in Fig. 6.3. We see that our findings indeed hold true for this depiction, and that the topmost accuracies in Fig. 6.2 are indeed dominated by results obtained on the data of pat04, and that the relative accuracies of different classification samples support our conclusions. It is to be noted specifically, that,

when taking a look not only at the distribution itself, but also the distribution of samples, we find quite a number of classification configurations, that are rightout outperforming every configuration achieved at all for QDA, AdaBoost or Passive Aggressive. This remark, small as it may sound, bears quite the significance, however, as, while we are currently investigating general suitability of different schemes and criteria for the characterization and classification of patients with depressive disorder, we are doing so to find classification schemes that are optimal for this kind of data, i.e. we are specifically interested in the configurations towards the top of these distributions.

Reviewing these accuracy distributions, we will, for the remainder of this analysis, focus on those schemes that yield a median accuracy of at least 0.85 over all participants included in this analysis, which means that the AdaBoost, the QDA as well as the Passive Aggressive classification schemes will be considered not sufficiently suited for the kind of data at hand, as they perform visibly worse compared to the rest of these algorithms.

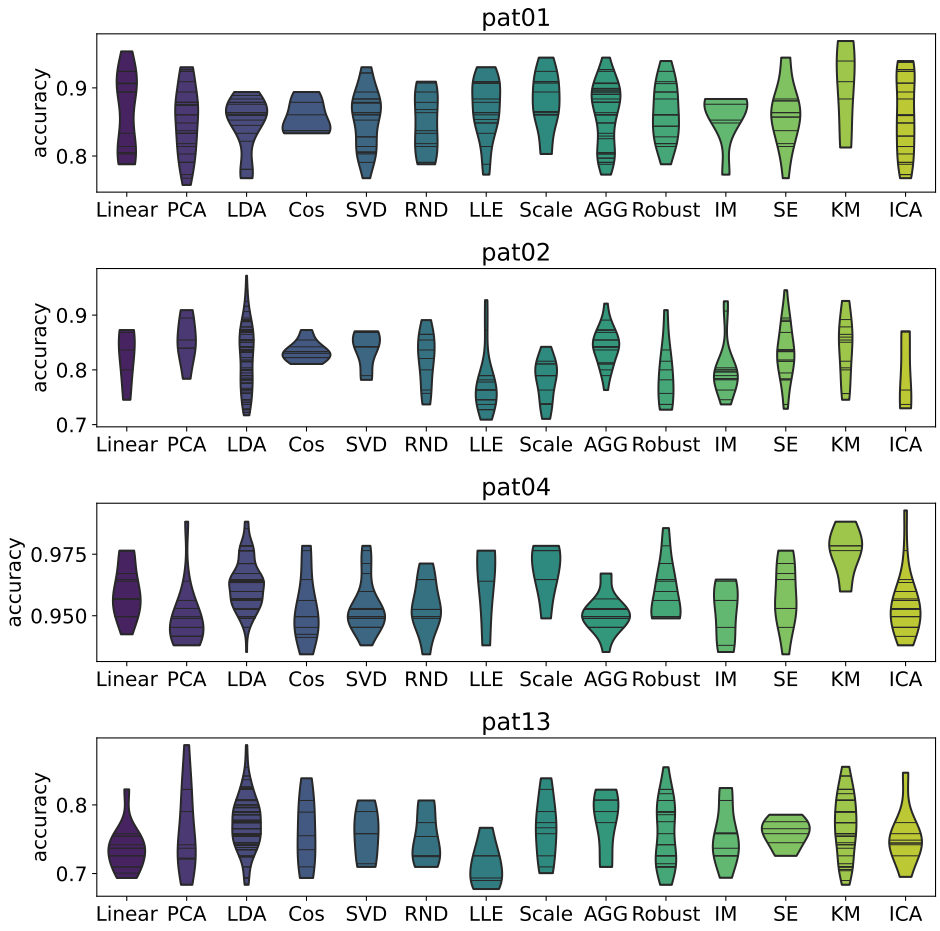

Figure 6.4.: Accuracy distribution associated with selected transformations split participant.

6.1.3. The impact of feature space transformations

While we could make a statement about the preferred classification algorithms, their accuracies will depend not only on the classification scheme itself, but also on the configuration of the underlying feature space. Hence, the suitability of a classification scheme depends on the consistency of the feature space used alongside it, as a classification scheme that achieves good classification performance with the same configuration over several participants will still be preferable in practice over a scheme that achieves slightly higher accuracies, yet with a fundamentally different configuration for every participant from an operational perspective. Hence, as a next step, we will investigate the contribution of the available transformations to achieving certain accuracy levels, again specifically for the different participants. Fig. 6.4 shows the distribution of transformations used in classification configurations with respect to their accuracy, the selection of the data-points is again the optima of each optimization procedure and the everything up to one percentage point below, with the horizontal stripes indicating the exact location of the associated configurations on the accuracy scale. A couple of things are immediately striking: Firstly, the sample density in the different transformation distributions in very unequal, already indicating that some are systematically preferred by the classification procedure. We will investigate this in more detail shortly.

Secondly, we notice that, while the exact shapes differ over participants, different transformations have different maximum accuracies that are achieved, confirming the hypothesis that transforming the dataset contributes to the achieved accuracy, especially compared to the linear transformation, which is just the identity, not changing the appearance of the dataset at all. In some cases, picking a different transformation achieves close to a 10 percentage points increase in accuracy, such as for pat13 as well as pat02, where we find that especially the LDA is capable to yield significant improvements in maximum achievable accuracy, also K-means does perform very well in all participants. We can furthermore compare the reductive transforms, namely the Linear, Robust, Scale and Cosinus transformations, with the dimensionality reducing transforms to conclude that a reduction in dimensionality of the dataset to be classified is not necessarily indicating an improved classification accuracy on its own. However, the transformations achieving best accuracy results are mostly dimensionality reducing, the ones with the least improvement in accuracy are as well. This strongly indicates that this effect is not a random artifact, but is a structural phenomenon. This is further supported by the fact that the Sparse Random Projection is not amongst the transformations correlated with top accuracies, but more alongside the lower end of the scale. Hence, dimensionality reduction on its own does not seem to be a recipe for improvement, rather the contrary. This indicates that not all inherent axes within the data contribute equally to the separability of the dataset, and that other transformations are more suited to extract those, such as the LDA. However, we also see that purely selecting components to condense to, such as it is done by PCA and ICA transformations, does also not guarantee improved accuracies. This can clearly be seen in the accuracy distributions for pat02, where the Sparse Random Projection performs approximately equal or better compared to PCA and ICA. The distributions for the dataset of pat04 is also indicative for this, where both PCA and ICA are capable of yielding absolute top accuracies, but the overwhelming majority of configurations transformed with these two mappings achieved lower performance compared to a Sparse Random Projection. A similar conclusion can be derived for SVD.

Thirdly, we notice an inhomegeneity of the selected transformations over the different participants. Noticeable examples of that include the LDA, which was able to achieve top or close-to-top accuracies for three participants, but, while still performing decent, capped out briefly below the 90% accuracy mark on pat01, whereas most accuracies managed to exceed that mark. A similar observation can be made for Local Linear Embeddings and pat13, however this could also be an artefact of the low amounts of samples that included this transformation in their configuration,

as for pat02, the LLE also mostly performed at the lower end of the scale with just a very few configurations achieving close-to-top performance.

Furthermore, we can draw conclusions about the structure of the underlying dataset given the performance of different transformations with respect to the optimization of separability. For one, we can observe that both Local Linear Embeddings and IsoMap, that both attempt to find structure within the dataset that is not represented by the standard representation, perform very similarly over all participants, are not amongst the top-accuracy yielding transformations, with the possible exception of LLE and pat01. This could be an indication that there is no inherent structure in the dataset that is hidden by its standard coordinate representation. That there *is*, however, a systemic underlying structure is hinted towards by the relatively good performance of cluster-based transformations, such as K-means as well as LDA, which both combined are associated with either the best, or second best accuracy achieved in every dataset, surpassed only by a configuration leveraging ICA as a transformation for pat04, but the accuracy improvement achieved by that configuration is extremely slim and also untypical for ICA transformed data for that dataset, raising the suspicion that it is more of an artifact than a consistent, reliable result. We do notice that the density of samples for K-means is comparatively sparse, especially for pat01 and pat04, but that doesn't come as a surprise given that the cluster separation that the Linear Discriminant Analysis is based on is performed by a K-means identification of clusters. Hence, the LDA, as used in this optimization procedure, could be understood as a K-means with an additional selection of axis for best separability. With this understanding, the consistently good performance of both LDA as well as K-means is a strong indicator that we do not only find cluster structures within the dataset representations, but also that we find more than just two clusters, because if we would have a clean separation of only two or very few clusters, we would expect transformations like PCA, ICA and SVD to be able to capture these structures as well, obtaining those axis during the dimensionality reduction that allow for separability of these structures. However, this is only partially the case, but especially with regard of the dataset for pat04 not a consistent finding.

Another aspect that can be seen in Fig. 6.4 is the different sample density for different transformations, as the selection of transform is part of the space open to the Bayesian optimizer. The accuracy itself is only part of the story, and Fig. 6.4 normalizes each distribution over the accuracies. Fig. 6.5, in contrast to that, depicts the number of samples histogrammed not with respect to their relative distribution, but in absolute numbers, underlining, what could already be assumed from assessing the previous figure, that especially the LDA is consistently chosen for several configurations. In this depiction, it is clearly evident that the LDA is not only used over the entire scale of achieved accuracy, which is especially noticeable for pat02, where the LDA consistently covers every accuracy bucket except one that is not even reached by any other except Spectral Embedding. Furthermore, we see that the numbers for all transformations except LDA are approximately evenly leveled in prevalence, whereas certain accuracy buckets show up to a 6-fold increase in prevalence for the LDA, mostly in the center area of the resulting accuracy distribution, but also towards the top, especially in the dataset of pat02. This even holds true for the pat01 dataset, where LDA was not amongst the transformations associated with the topmost accuracies, but still consistently outpaced all other transformations when it comes to prevalence. This clearly makes it the most versatile, and consequentially the most robust choice of the available transformations for a variety of configurations, and hence attractive for the classification configuration applied to in practice for patient treatment.

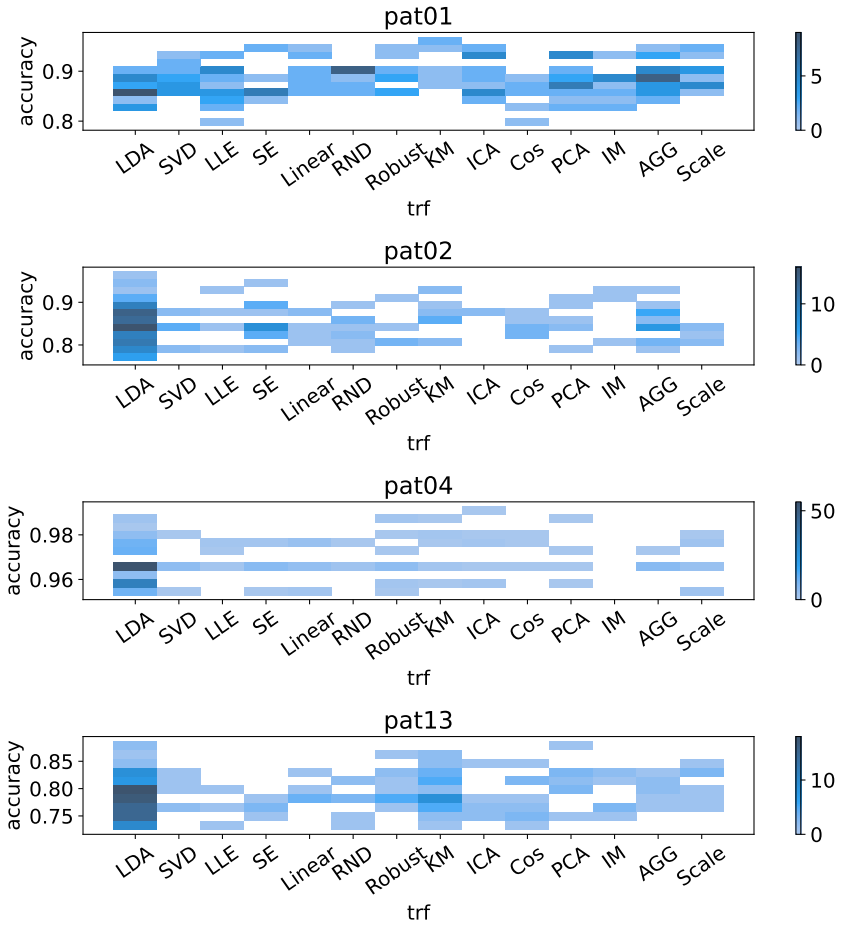

Figure 6.5.: Accuracy distribution histogram of prevalence of selected transformations in the optimization procedure, split by participant.

6.1.4. Investigating the interdependence of classification schemes and feature space transformations

However, so far the analysis of the suitability of transformations fell short in one aspect: While the up to 6-fold increase in prevalence clearly indicates the likelihood that LDA performs consistently well over several algorithms, this is still unverified. Hence, the direct next step is verifying this not being an artefact of a single algorithm. To do so, we will investigate the distribution of transformations not only split by participant, but also by algorithm, as depicted in Fig. 6.6. Several notable things can be found in this depiction: We indeed do find a consistent use of LDA over several algorithms and several participants, with two notable exceptions: Firstly we find that LDA does not hold its high prevalence in the dataset for pat01. This is consistent with the finding in 6.4, where LDA did not yield the topmost accuracy or close to it, whereas PCA, SVD and ICA as well as geometric embeddings were significantly more prominent. Furthermore, we see that in three out of four datasets, the Naive Bayes is also very unassociated with the LDA. Similar observations can be made for tree-based classifiers, although that observation is not as

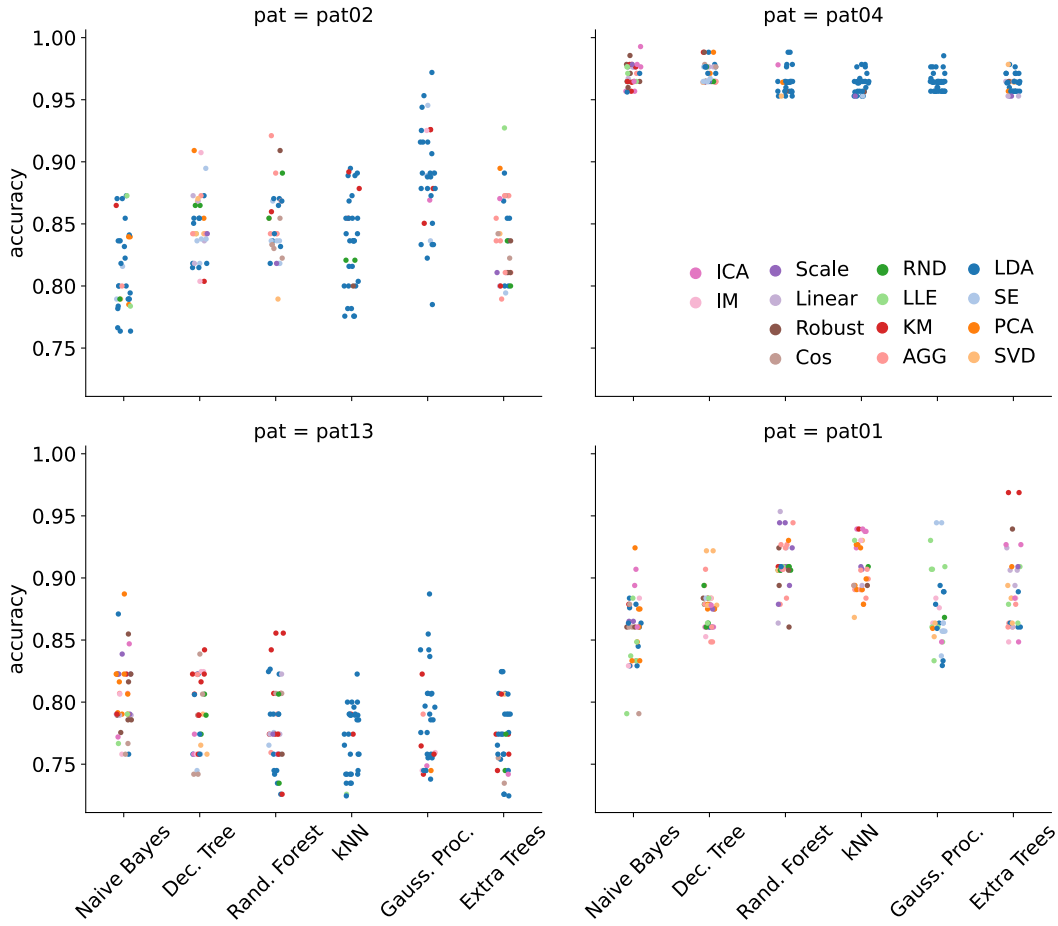

Figure 6.6.: Accuracies associated with algorithms, transformations and participant. Each point depicts the accuracy of a resulting configuration, split by algorithm (column) and feature space transformation applied (color) per participant.

strong as with the Naive Bayes, indicating that the geometric nature emphasised by the choice of dimensions by the LDA is not as relevant in this particular case. This is further emphasised by a higher-than-average prevalence of non-dimensionality reducing transformations, indicating that the Naive Bayes classifier doesn't substantially benefit as much from dimensionality reduction of the search space as other classification schemes do. In fact, we do not see a substantial difference in accuracy between reductive and nonreductive transformations in the distribution of configurations for the Naive Bayes. This implicitly makes it, to a degree, a benchmark of how effective dimensionality reduction techniques work, and we can conclude that, while they are clearly beneficial to some classification schemes, they are not strictly necessary to achieve good data partition. However, we find that especially the LDA is capable of aiding constant classification accuracies for different configurations. This is also true of other transformations, but especially pronounced for the LDA, as it can be observed to create bands on multiple accuracy levels, indicating that it is capable of partitioning the dataset equally for different configurations of data fields. This, for one, underlines the effectiveness of LDA in stabilizing classification

accuracy as well as underlining the consistent partitioning of data for different configurations. This can also be observed for the pat04 dataset for the Naive Bayes scheme, which shows this band behavior as well with a mixture of transformations, making this property nonexclusive to the LDA, however the LDA is most prominently exhibiting it for all tree-based classifiers, k-Nearest-Neighbours as well as Gaussian Processes and even for the Naive Bayes for the pat02 dataset. We also find again that the preferred classifier is not consistently the same for the different datasets, but that the preferred classifier varies in each dataset. Gaussian Processes are consistently high for multiple datasets, even forming a stabilized band in combination with the LDA in the pat02 dataset, which is in contrast to the other datasets dominated by non-LDA transformations, and the Random Forest classifier also performs overall consistently well. This could be yet another indication that it would be advisable to redo this analysis in detail with a greater amount of acquired data, which we at this point in time do not have at our disposal.

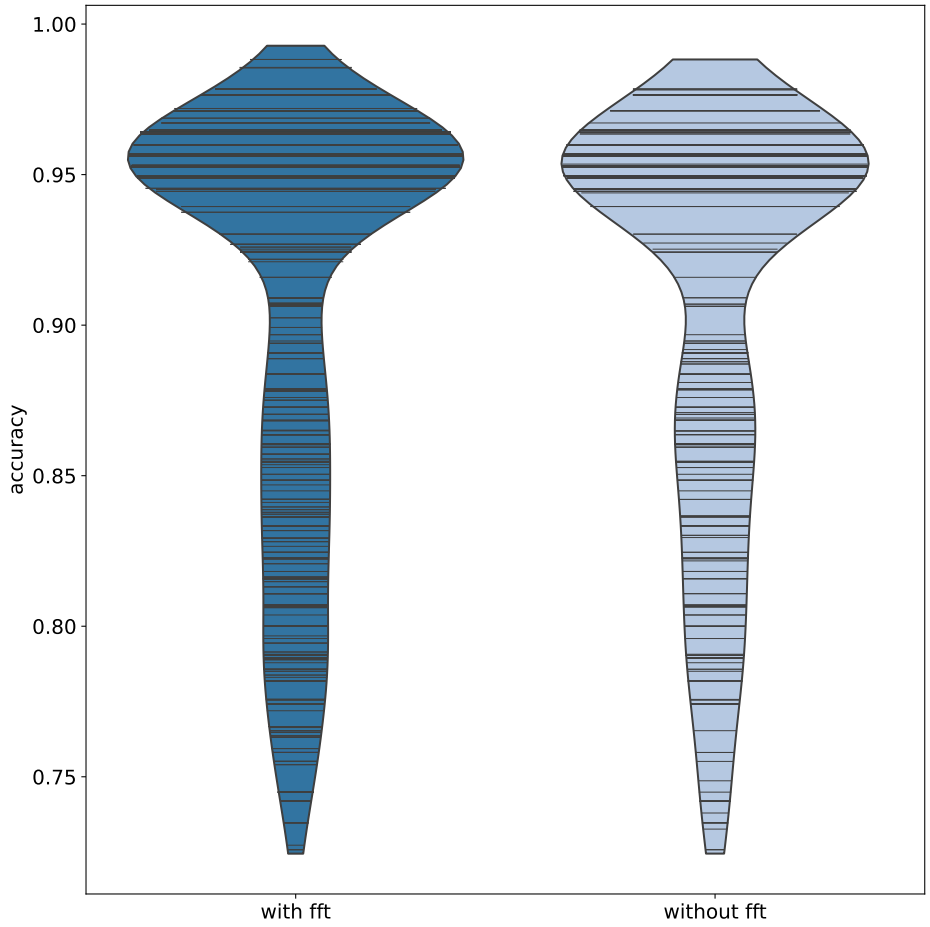

Figure 6.7.: Accuracy distribution of configurations using Fourier-transformed data and not using Fourier-transformed data. Classification has been performed participant-individualistically before unionizing for display.

6.1.5. The effect of frequency bands

The next target of analysis will be the effectiveness of the Fourier transforms to the classification. Fourier fields are typically used to complement the non-transformed data, and to provide insight into potential information in frequency bands. To investigate its effectiveness, and as a consequence the hypothesis that participants in different states exhibit characteristic features that show primarily in such frequency bands, but cannot be seen in other kinds of data, we can take a look at Fig. 6.7, which takes the same data selection criterion we have used so far, but distinguishes between configurations that include frequency band data in some form and those that do not include frequency data at all. The result is as straightforward as simple: there is almost no change in distribution of accuracies, strongly indicating that the frequency band information, while potentially useful in some configurations, can be entirely replaced by configurations that do not rely on it at all or are just marginally contributing to improved accuracy, potentially being a statistical artifact. This is also emphasised when we split configurations data fields and

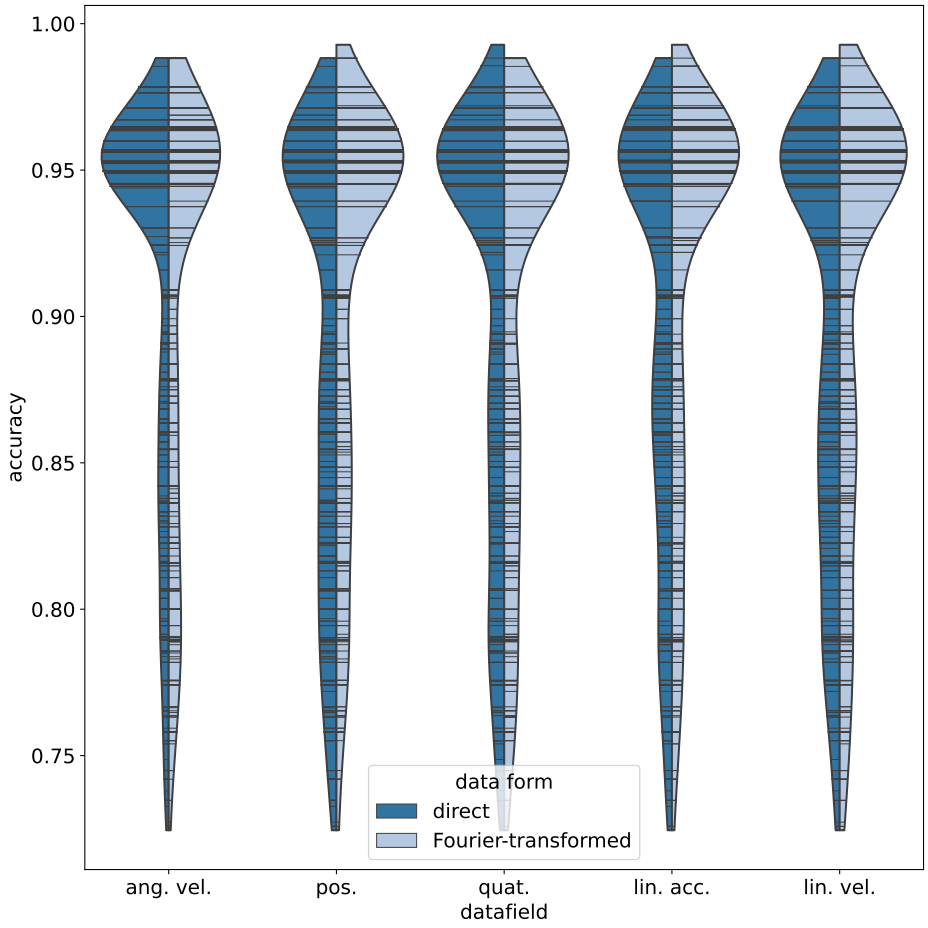

Figure 6.8.: Accuracy distribution associated with different base data types. Classification has been performed participant-individualistically before unionizing for display.

investigate their association with accuracy individually, as seen in Fig. 6.8. Here, we also do

not see any improvement yielded by frequency band information, leading us to the consequential conclusion that frequency bands to not contain any information that is of substantial use for the goal of distinguishing between dysfunctional and non-dysfunctional mental states in patients diagnosed with depressive disorder.

6.1.6. Deconstructing accuracy measurements

Up to now, we have focused purely on the absolute accuracy of the obtained results. However, different factors can contribute to such an accuracy, which is what we will investigate next. To do so, we will take the accuracy results that we have investigated up to this point and instead of looking at the accuracy numbers, we will instead take a look at three different properties that allow us to better understand how these accuracies emerge.

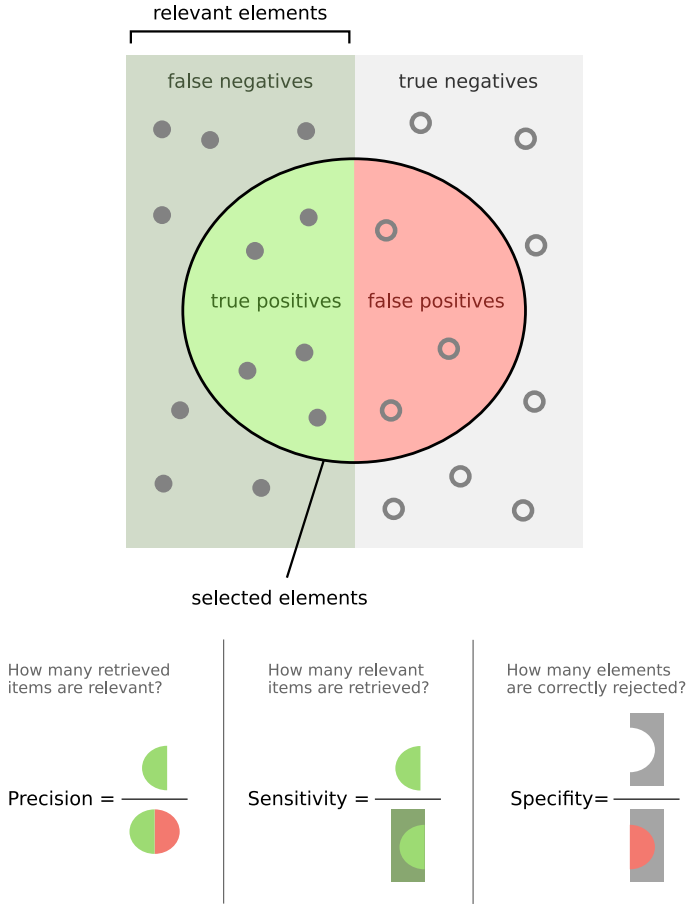

Figure 6.9.: Illustration of precision, sensitivity and specificity, adapted from [95].

To do so, we first define the definitions

$$tp := \textit{true positive, a correctly classified dysfunctional sample} \tag{6.2}$$
$$tn := \textit{true negative, a correctly classified non-dysfunctional sample} \tag{6.3}$$
$$fp := \textit{false positive, a non-dysfunctional sample classified as dysfunctional} \tag{6.4}$$
$$fn := \textit{false negative, a dysfunctional sample classified as non-dysfunctional} \tag{6.5}$$

with which the accuracy a of a classification result reads

$$a = \frac{tp + tn}{tp + tn + fp + fn} \ . \tag{6.6}$$

With these definition, we can now define further properties to gain a deeper understanding how the described accuracies emerge.

Sensitivity

The properties we will look at are, firstly, the sensitivity to or recall of dysfunctional events, i.e. the amount of dysfunctional events that are correctly recalled among all dysfunctional events.

Formally, sensitivity θ reads, as also illustrated in Fig. 6.9,

$$\theta = \frac{tp}{tp + fn} \ . \tag{6.7}$$

Sensitivity and recall are two terms for the same property, the former originating in medical science whereas the latter originates in the field of pattern recognition. Due to the medical application of this work, we will use the term sensitivity for the remainder of this thesis. The results, split by participant and algorithm, can be seen in Fig. 6.10.

The results provide a clear picture: we find that accuracy is not generally strongly dominated by sensitivity, indicating that in future applications, if sensitivity is chosen to be a target property, it should be directly incorporated into the objective function for the optimization procedure. Furthermore, however, we do find that sensitivities can actually be extremely high, indicating that, if incorporated into the objective function, sensitivities can be a viable optimization target for both this procedure as well as the data at hand. We also see that the level of correlation of accuracy and sensitivity varies between datasets as well as between algorithms. A general observation is that both the k-Nearest-Neighbours scheme as well as the Gaussian Processes have a tendency to lean towards higher sensitivities for high accuracies, which is especially pronounced in the pat13 dataset, where both algorithms show sensitivities with a mode of around 0.8 and a number of samples achieving even higher sensitivities. This is less pronounced in the pat04 dataset, but still both the kNN as well as the Gaussian Process are found to have a substantial number of configurations that achieve very high sensitivities. This is especially noticeable compared to the decision tree and Naive Bayes algorithms, which exhibit extremely poor sensitivities for these high accuracies compared to other algorithms in all datasets but the pat01 one, where the Naive Bayes exhibits the most favorable distribution. Interestingly, it can also be observed that this favorable sensitivity distribution does not translate to a similarly favorable distribution in accuracy, in fact the accuracy performance of the Naive Bayes on the pat01 dataset is rather average compared to the other algorithms involved, whereas the sensitivity performance on the dataset that actually slightly favored the Naive Bayes with regards to accuracy, the pat13 dataset, yields the arguably worst distribution of sensitivities, further stressing the point that a good accuracy is not necessarily based on a good recognition of actual dysfunctional states. Furthermore, we see a clear tendency over all datasets that tree-based methods underperform

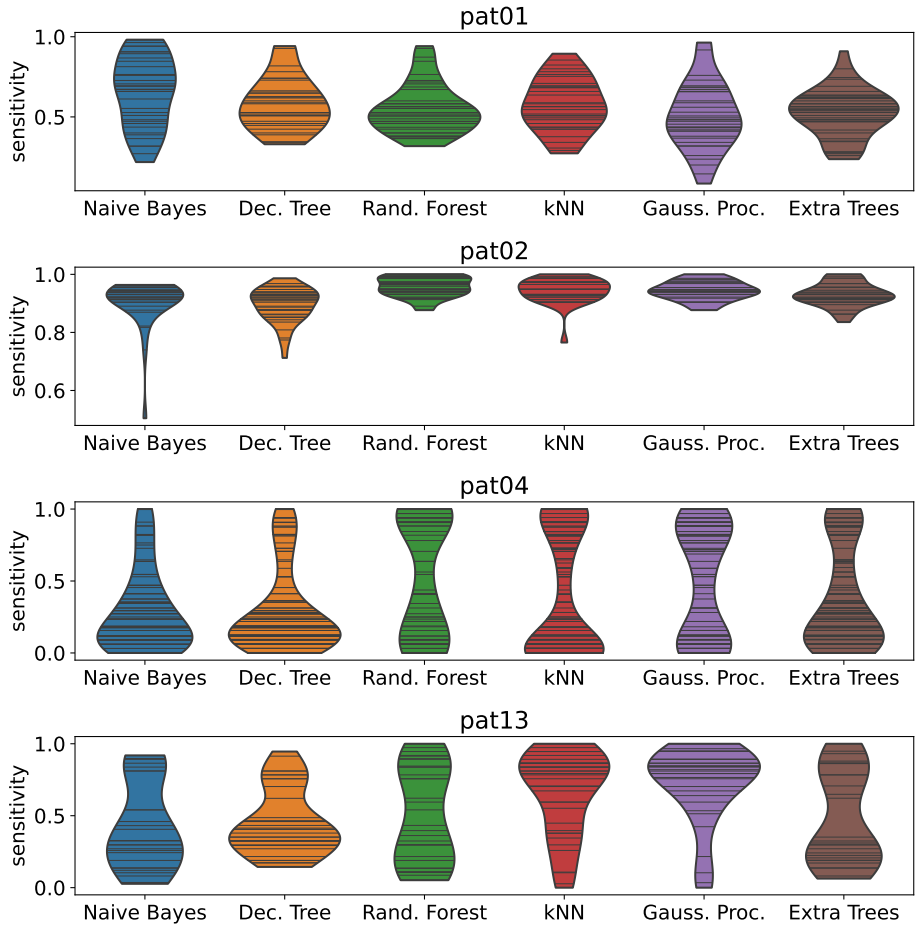

Figure 6.10.: Sensitivity distribution associated with high accuracies.

here, as high sensitivities are not that correlated to high accuracies for these types of algorithms. This could point to a systematic weakness in tree-based methods, because even if sensitivity is a desired objective, overall accuracy will also be desirable to avoid overdetection of dysfunctional states where there are none, which could lead to the risk of patients distrusting such a technology, assuming it just regularly triggers. Hence, a high correlation of accuracy and sensitivity is a very desirable property for such a setup, favoring especially the k-Nearest-Neighbours and the Gaussian Processes, all datasets considered. That being said, we still see that sensitivities of appropriate value can be obtained, as all distributions reach top sensitivities very close to 1, meaning that all algorithms are capable of actually detecting all dysfunctional states that were part of our test data.

Specificity

Besides sensitivity we can investigate the contribution of specificity, which is in a sense the inverse of the sensitivity, to the accuracy distribution. Specificity describes how many non-relevant events were classified as such, i.e. where the sensitivity describes the fraction of relevant

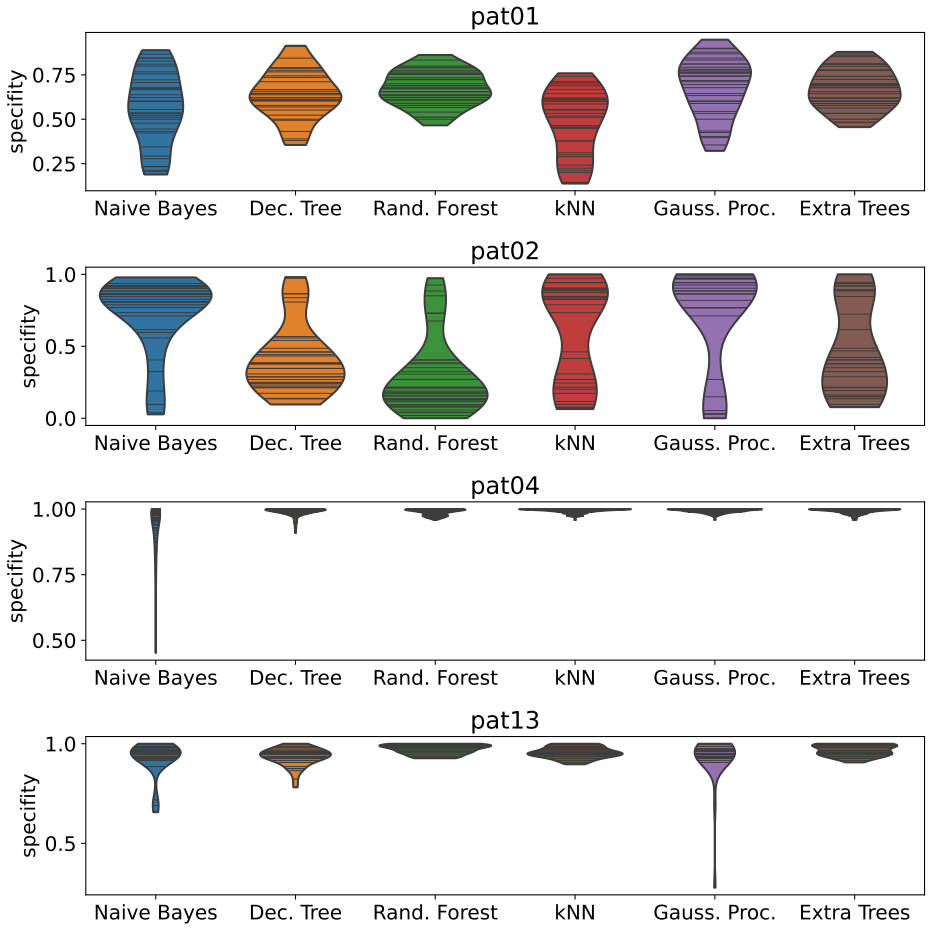

Figure 6.11.: Specificity distribution associated with high accuracies.

samples that are detected, the specificity is a measure of how much the classification procedure does not overclassify (again, see Fig. 6.9 for a more detailed illustration). Formally, we will define specificity γ as

$$\gamma := \frac{tn}{fp + tn} \ , \tag{6.8}$$

i.e. the fraction of correctly classified non-dysfunctional states amongst all non-dysfunctional samples, or the true negative rate.

Investigating the specificity distribution of the set of selected configurations per algorithm, analogous to the investigation of sensitivity, we find a number of interesting aspects, which can be seen in Fig. 6.11. Firstly, we find an overall much stronger correlation between accuracy and specificity, which is a strong indicator that non-dysfunctional states are easier to track compared to dysfunctional states. This seems to be especially pronounced in pat13, where most of the configurations achieve a specificity of about 90% and higher, with the few exceptions being the Naive Bayes and the Decision Tree, as well as some extremely bad configurations of the Gaussian Process, but these are very rare, so overall the Gaussian Process seems to be able to systematically deliver specificities around and above 90% for this dataset.

The specificity exhibited on dataset pat04 is even better, but here we need to recall that this dataset is very skewed towards non-dysfunctional states, limiting the reliability of the specificity estimates on this one due to the naturally emerging lower limit on the specificity at high accuracies, as the accuracy cannot be achieved by correctly detecting dysfunctional states alone due to their limited prevalence in the dataset.

That being said, we see two effects immediate effects: First of all, we see that again, k-Nearest-Neighbours and Gaussian Processes are the ones achieving high specificities most consistently, witch is most pronounced for the datasets pat02 and pat04, kNN performs slightly below average on the pat01 dataset but the Gaussian Process scheme achieves highest specificities most consistently, with a strong band around 80% and a number of samples above that. on the pat02 dataset, the most pronounced band is even higher, forming close to absolute specificity, whereas the kNN band is slightly below that. The Naive Bayes performs worse than expected: while its specificity still highly correlated with overall accuracy in the dataset pat02, this is less so the case in the other three datasets, placing it on the lower end of achieved specificities. This includes a rare occurrence of specificities of about 50% in the pat04 dataset, which must be assumed to be a numerical artifact. Interestingly, we find that tree-based algorithms, which performed either average or particularly sub-par in the sensitivity analysis, do so again in the specificity analysis, performing particularly unfavorable on the pat02 dataset, showing no strong-suit in either.

Considering both the results of the sensitivity as well as the specificity distributions, several things can be concluded: Accuracy serves as a comparatively good predictor for both sensitivity and specificity particularly for the Gaussian Processes and k-Nearest-Neighbours schemes, which have been shown to be able to produce configurations of high sensitivity as well as configurations of high specificity when trained for accuracy, which makes them extremely versatile. In addition to this especially the k-Nearest-Neighbours is conceptually very simple, therefore being a good choice given the regulatory proceedings that require explainability of the derivation of predictions. Tree-based methods, on the other hand, while performing well in single instances, do not show that their accuracy estimates are systematically correlated with either sensitivity or specificity, as the distributions are disproportionally accumulating most of the investigated samples at the bottom end of the sensitivity as well as the specificity distribution, performing worse compared to Gaussian Processes or k-Nearest-Neighbour schemes. Something similar can be observed for the Naive Bayes classifier, whose high accuracies seem to be disproportionally achieved by high specificity results, whereas it performs either average (pat01 and pat02 datasets) or amongst the rear end of the field (pat04 and pat13 datasets) in the sensitivity analysis, while achieving fairly good results in at least one specificity analysis (pat02 dataset). Given that modifications of the Naive Bayes procedure such as QDA have been substantially underperforming compared to the rest of the field, it does not seem likely that further modifications to its probability distribution function will noticeably improve its results. Scheme-agnostically, the results overall indicate that non-dysfunctional states are less difficult to assess than dysfunctional states, which is both plausible, as the non-dysfunctional states should not only be the norm of human motion, but are also characterized by higher overall motion in general [81]. However, this also underlines the challenge we face in our objective, as the self-reported dysfunctional states are those that we intend to primarily capture to act upon.

Precision

We will now shift our analysis to further assess which part of the accuracy is contributed by correctly assessing dysfunctional states and which are contributing by correctly assessing non-dysfunctional ones. To do so, we will investigate the distribution of precision ρ of these configurations, which describes the fraction of correctly assessed dysfunctional states amongst all states

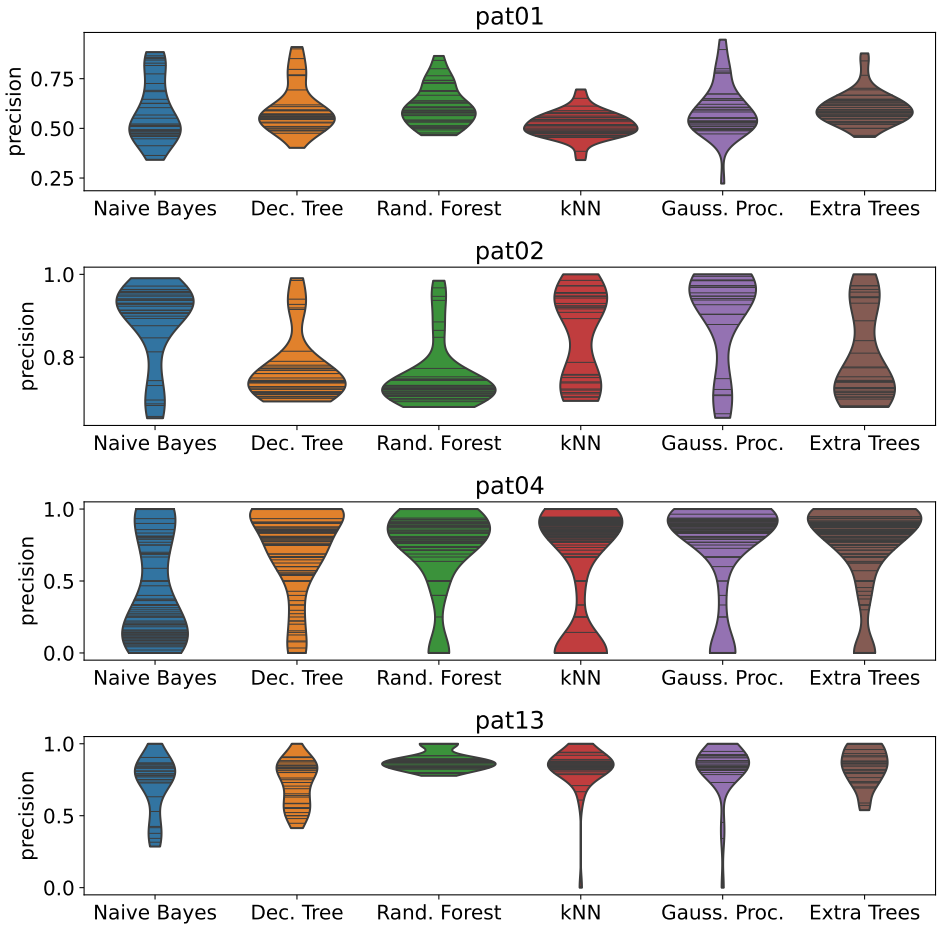

Figure 6.12.: Precision distribution associated with high accuracies.

identified as dysfunctional, i.e.

$$\rho = \frac{tp}{tp + fp} \ , \tag{6.9}$$

the results of which can be seen in Fig. 6.12. For an illustration of the term in this context we again refer to Fig. 6.9. The results we find are conceptually similar to the accuracy distribution itself, but here we only consider states of interest, i.e. dysfunctional samples, for the accuracy estimate. If the accuracy is dominated purely by the precision of a classification configuration, we would expect the precision distribution to yield higher values than the accuracy distribution, whereas we would expect a precision below the accuracy for configurations that are primarily detecting non-dysfunctional states. Comparing the accuracy distribution with the precision distributions, a number of core observations are immediately striking: Firstly, we do find the precision distribution of the pat04 dataset to reach extremely low, some configurations are bottoming out at plain zero. This is, however, unsurprising as this is the dataset that is highly skewed towards non-dysfunctional states. Consequentially, it is very simple to reach high accuracies with low precision on this particular dataset. We, however, also see, that the precision is still substantial for most algorithms, with the most dense accumulation of configurations ranging

solidly around or above the 90% mark, indicating that most configurations are still able to precisely explicitly select dysfunctional states, with fairly few false positives only. The one exception to this rule is the Naive Bayes, which is in line with its very suboptimal performance sensitivity-wise on this dataset, indicating that the Naive Bayes in this dataset exhibits general difficulties with correctly detecting dysfunctional states. All other algorithms exhibit a top-heavy precision distribution indicating suitability for the correct detection of dysfunctional mental states, with the decision tree scheme performing slightly less top-heavy compared to the rest. Focusing on dataset pat13, we can see that, while the precision distribution tailors downwards quite a bit for some configurations, the majority of configurations actually achieves a better precision compared to the accuracy distribution, indicating that accuracy is actually driven by the correct classification of dysfunctional states. This is the case for the Gaussian Process, k-Nearest-Neighbours, Extra-Tree and Random-Forest schemes, whereas the Naive Bayes and the Decision Tree show a distribution that, while capable of exceeding the accuracy distribution and hence indicating that there are in fact configurations whose accuracy is driven by correctly classified dysfunctional states, exhibits also configurations that are lower than the accuracy distribution, hence not showing such a clear indication which state actually drives the classification accuracy. This result yields an interesting combination when it is combined with consideration of the dataset pat02, where we find a different distribution of precisions compared to accuracy: here, the k-Nearest-Neighbours as well as the Gaussian Processes yield precision distributions above the corresponding accuracy distributions, showing that in this dataset the classification accuracy is also driven by the correct detection of dysfunctional states, the Gaussian Processes even more so compared to the k-Nearest-Neighbours. We, however, also observe that tree-based schemes yield underperforming precisions compared to the corresponding accuracy distribution, specifically the Random Forest and Decision Tree methods, and while to a lesser degree, this is also the case for the Extra Trees scheme as well. This finding is mostly unsupported by the sensitivity results we find for these classification schemes, where only the Decision Tree scheme performed slightly worse, whereas the Random Forest and Extra Trees schemes performed about on par with the k-Nearest-Neighbours and Gaussian Process schemes. This is no longer the case in the precision evaluation, indicating a higher tendency to achieve its sensitivity by overdetection. The Naive Bayes, on the other hand, yields surprisingly good precisions for a slightly underperforming sensitivity, indicating that while it is generally performing sub-par in detecting dysfunctional states, it at least does so fairly reliable. These findings are not as pronounced in the dataset pat01, in which all classification schemes are driven primarily by the detection of non-dysfunctional states, whereas the resulting precisions can merely be described as mediocre at best.

6.1.7. Concluding Interpatient analysis

The results so far indicate a number of substantial findings: Firstly, we find that the performance of a configuration and also a configuration option does depend significantly on the dataset. While unsurprising, this still quantitatively substantiates the high level of individualism that needs to be considered when describing depressive disorder, and it serves as an indication that such a system will have to be tuned to the specific kind of motion that is applicable to a specific patient. Furthermore, the results seem to be inconsistent over different participants when it comes to not only accuracies, but the breakdown of these accuracies into sensitivity, specificity and precision, creating the impression that these are also highly patient specific, which is a corollary from the previous point. However, we do find consistency in the inconsistency when we compare the prevalence of transformations in the high accuracy set with, exemplarily, the precision behavior of different classification schemes. Ignoring the highly skewed and to a degree borderline pathological label distribution of dataset pat04 for a precision driven accuracy, we still find configurations that perform above-accuracy in terms of precision, being consistent with

what we see in the precision distribution of datasets pat02 and pat13. The outlier here is clearly and systemically in all schemes the pat01 dataset, and this is consistent with its state as the outlier when it comes to transformations, being the only dataset that barely contains any LDA-transformations in its high-accuracy dataset. This indicates the pat01 dataset to be somewhat untypical compared to the other three datasets that have been evaluated thus far, highlighting both the need for individual configurations, but also indicating that such a procedure, when generalized, will not necessarily be suited for every patient in therapy, but might instead need to be evaluated on its effectiveness during therapy. For an effective therapeutical application of such a measure, it might therefore be advisable to either test different configurations or evaluate the option of ensemble classifiers with higher degrees of variability in configuration.

That being said, as we do find consistencies in configuration between three of four datasets under investigation thus far, it seems likely that some participants do share general patterns of motions to be indicative for dysfunctional and non-dysfunctional states. Hence, the question arises if we can supplement patient-specific measurements with a general baseline of previously acquired patient measurements to increase the usefulness of such a method in therapy by providing quicker convergence. Therefore, in the next section we will discuss the option of data mixing between participants and its consequences.

6.2. Mixing Data: a cross-sectional study

To investigate cross-participant results for our procedure, we first need to define a baseline dataset. To do so, we will do three things:

- lower the amount of data available for inclusion to 3 hours
- increase the number of participants given this reduced inclusion criterion
- stratify both the amount of data taken per participant as well as the amount data taken per label per participant

The first and second criterion represent a tradeoff made to diversify the resulting dataset, given that only four participants provided enough data during the study to meet the inclusion criterion of the first part of this study, while still retaining an appropriate size of the resulting dataset. The last point is substantial to avoid the optimization routine to be dominated by those participants who contributed the most data. This did not pose an issue in the first part of the study, where participant data was analysed separately per participant, but when their data is combined into one dataset, a significant imbalance would skew the result towards the dominating data subset, therefore the data taken per participant was capped at three hours, and was selected in a way that ensured that each participant would contribute approximately equally to both data subsets representing dysfunctional and non-dysfunctional states to avoid the described effect occurring should one participant be overrepresented one and underrepresented in the other data subset. This stratification process was applied to both the training set that was used to perform the optimization as well as the testset on which the resulting accuracy figures were estimated.

6.2.1. Algorithm-specific accuracies

Using this dataset, we firstly want to investigate the accuracy per algorithm, to verify that it still performs better than just random guessing, which in this dataset would come out at 50% accuracy, and furthermore compare this to the accuracies achieved on the individual datasets. The results, which are depicted in Fig. 6.13 and again obtained as a collection of results of multiple optimization runs with fixed classifier and taking each run's optima including the results

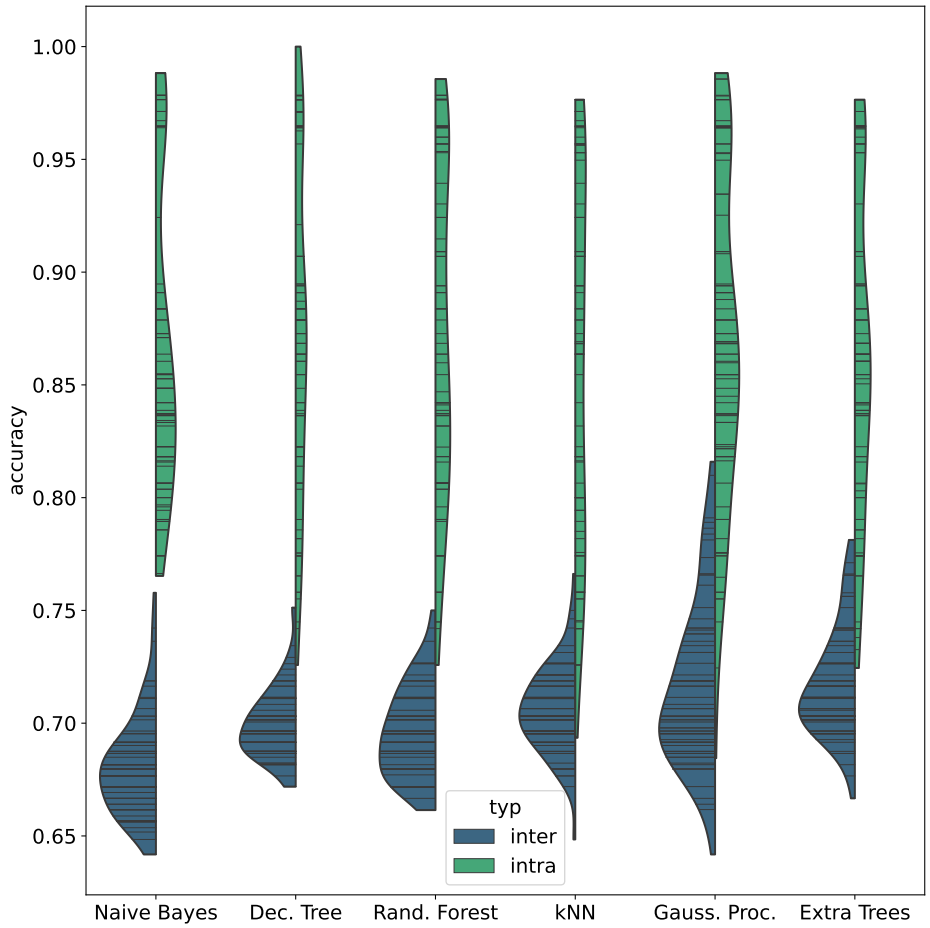

Figure 6.13.: Accuracy distribution of inter- and intra-participant analysis. Intra-participant data obtained as previously discussed and unionized for display.

yielding accuracies up to a percentage point lower than the run's optimum, clearly show two things: Firstly, we do see that the resulting accuracies are not substantially, but significantly above 50%, accumulating samples around up to 67% to 70% accuracy, with the Gaussian Process being capable of reaching up to 80% accuracy and the Extra Tree scheme reaching up to 77%. The other tree-based schemes as well as the kNN yield results accumulating around 70%, with the Naive Bayes again indicating that it is very hit-or-miss, coming out worst of all algorithms considered at around 68%. This shows that even on this more difficult dataset the procedure is capable of yielding substantial improvements in accuracy, which we can use to derive properties of the underlying dataset. Secondly, the dataset also shows a substantially worse performance compared to the individual datasets, which have their individual distributions merged into one for simpler comparison. This is unsurprising, given that the individual analysis already indicated that motions indicative for dysfunctional depressive states are highly individual. Hence, finding the one-fits-all solution did already seem difficult in the last section, and this seems to be confirmed here. However, it still serves both as a baseline of what is achievable, stressing the substantial improvement of the results presented in the last section as well as hinting at

some degree of systematic structure between states in the data, but also serving as a baseline to improve upon with individual data of a participant using the system, as well as an indication of what kind of setup is capable of yielding these kinds of accuracies in such a mixed dataset.

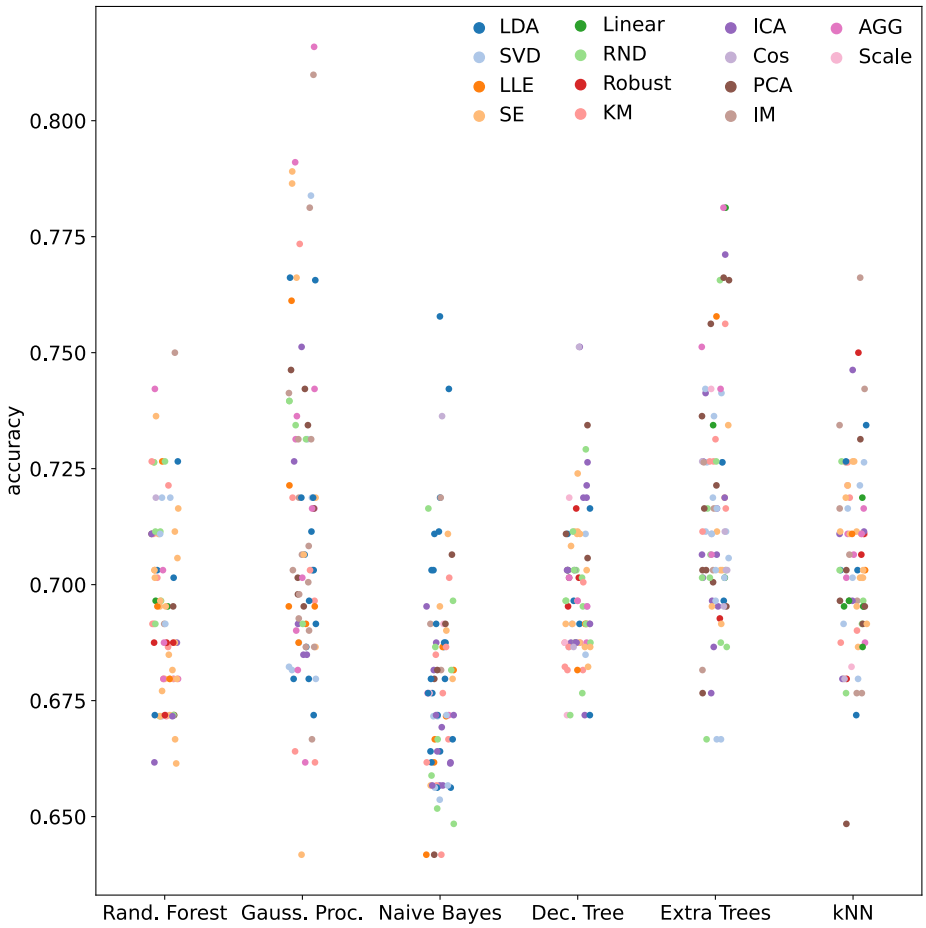

Figure 6.14.: Accuracy distribution of inter-participant analysis split by algorithm and transformation.

6.2.2. The impact of feature space transformations

One of the aspects we are revisiting in the analysis is the distribution of transformations over both classification schemes and accuracy, which for the interparticipant dataset is depicted in Fig. 6.14. We observe a number of things, the most prominent being that in the interparticipant dataset, the LDA, which was the most used transformations in the accuracy, is substantially less prominent, still being most prominently combined with the Gaussian Process as well as the Naive Bayes. Furthermore, we observe less of a definite preference of the optimization procedure towards one specific transformation. This is a strong indication of what we already discussed, namely that the dataset at hand is now significantly more complex. While so far this could be observed in reduced accuracies, it is now also observable in the less stringent choice of transformations. As we, in contrast to the participant-specific datasets, can only observe marginal

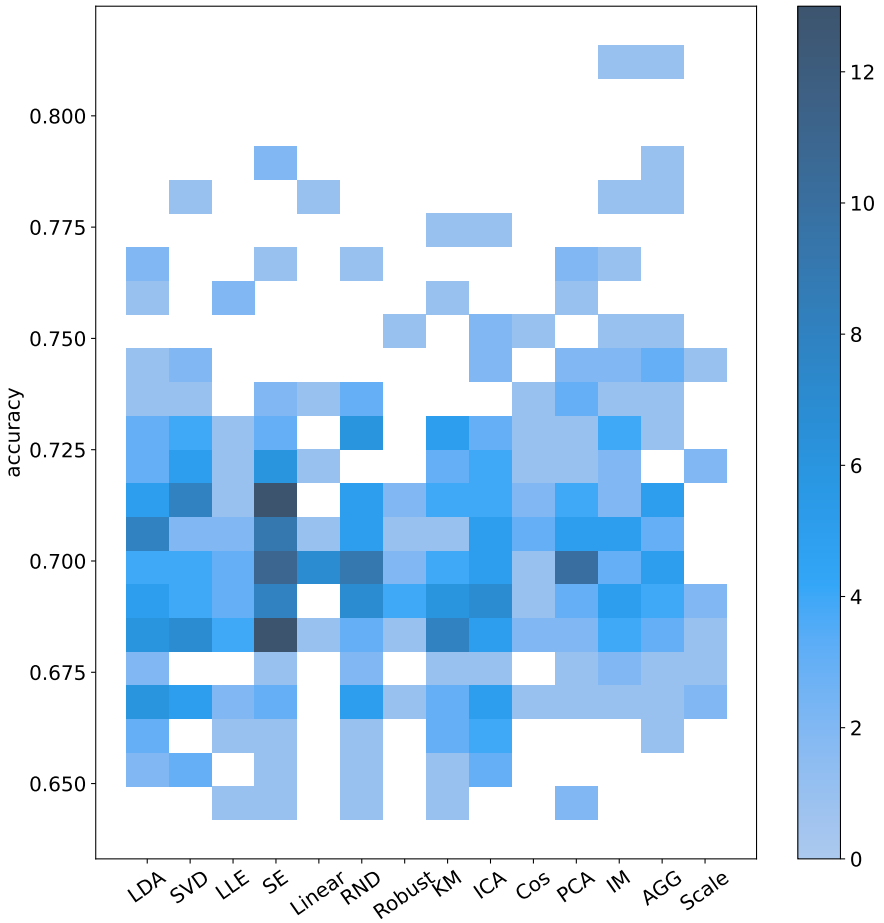

Figure 6.15.: Prevalence histogram of accuracy distribution associated with different transformations on the inter-participant dataset.

synergies between specific algorithms and transforms, it seems more advisable to investigate accuracies for all algorithms combined, assessing which transformations are associated with which accuracies, which is depicted in Fig. 6.15. Here we find a substantially more even distribution over the different transforms, again clearly reinforcing what could be deduced from the results depicted in Fig. 6.14. We find a barycenter of accuracy around the 70% mark and we also find a very even distribution over the different transformations, with configurations achieving accuracies beyond 75% becoming substantially sparse. In the barycenter area of the accuracies associated with different transformations, we can also observe some key points: Firstly, the non-reductive transforms like Linear, Robust and Scale transformations are noticeably sparser in their usage compared to other transformations, strongly indicating that in this dataset in particular, a lot of dimensions are not relevant for appropriately classifying samples, hence a reduction is not only possible but seems to be actually helpful in improving classification performance for most configurations. In contrast to that, while the LDA finds about average usage on this dataset, we find a substantial tendency of the optimization routine towards a Spectral Embedding, with also Principal Component based as well as a Single Value Decomposition based reductions finding more

common usage. However, the histogram is overall comparatively flat, making it unlikely that this is a significant finding. This scepticism is also supported by taking a look at other Embedding techniques, such as IsoMap and Local Linear Embeddings. If the usage of Spectral Embeddings would point towards a relevant geometric substructure that is not captured by the representation of the data, we would expect to find a focus on the embedding-based transformations instead of a more broader usage of different transformations. Also, the Sparse Random Projection is used more prominently compared to the participant-specific analysis, which was heavily dominated by LDA, indicating that complexity of the resulting data is a hindrance to proper classification. This is very likely the reason why SVD and PCA are found with high prevalence as well, as they do the same procedure, albeit in a more directed fashion. The same is still true for LDA, choosing a advantageous subspace for classification, even if its specific procedure does not exhibit an effectiveness comparable to the participant-specific analysis.

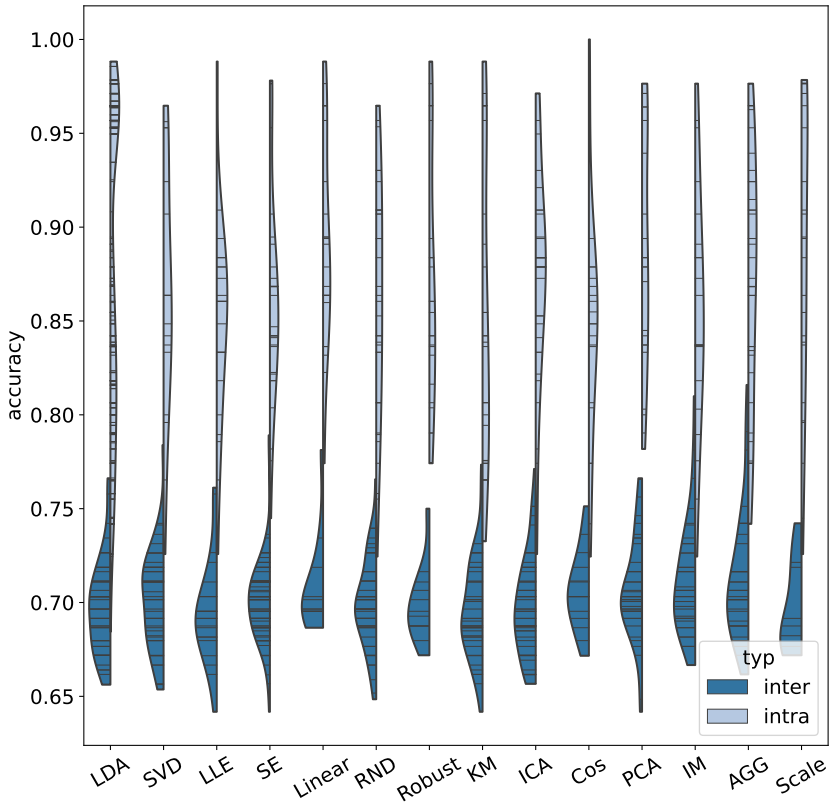

Figure 6.16.: Accuracy distribution of inter- and intra-participant analysis split by associated transformation. Intra-participant data obtained as previously discussed and unionized for display.

We will briefly touch on the comparison of effectiveness in direct comparison to the participant-specific analysis, which is depicted in Fig. 6.16 and effectively, given that we do not find a transformation that shows particularly remarkable properties, replicates the findings we discussed

when comparing the accuracy of algorithms: We do find an expected regression in accuracy, and we do not find substantial differences between different transformations besides what has already been discussed in the previous paragraph. For comparison, this data is shown against the union of patient-specific results, which however does not yield any additional information. Still, equal accuracy does not imply equality when it comes to sensitivity and specificity, which we will investigate next.

6.2.3. Deconstructing accuracy emergence by feature space transformation

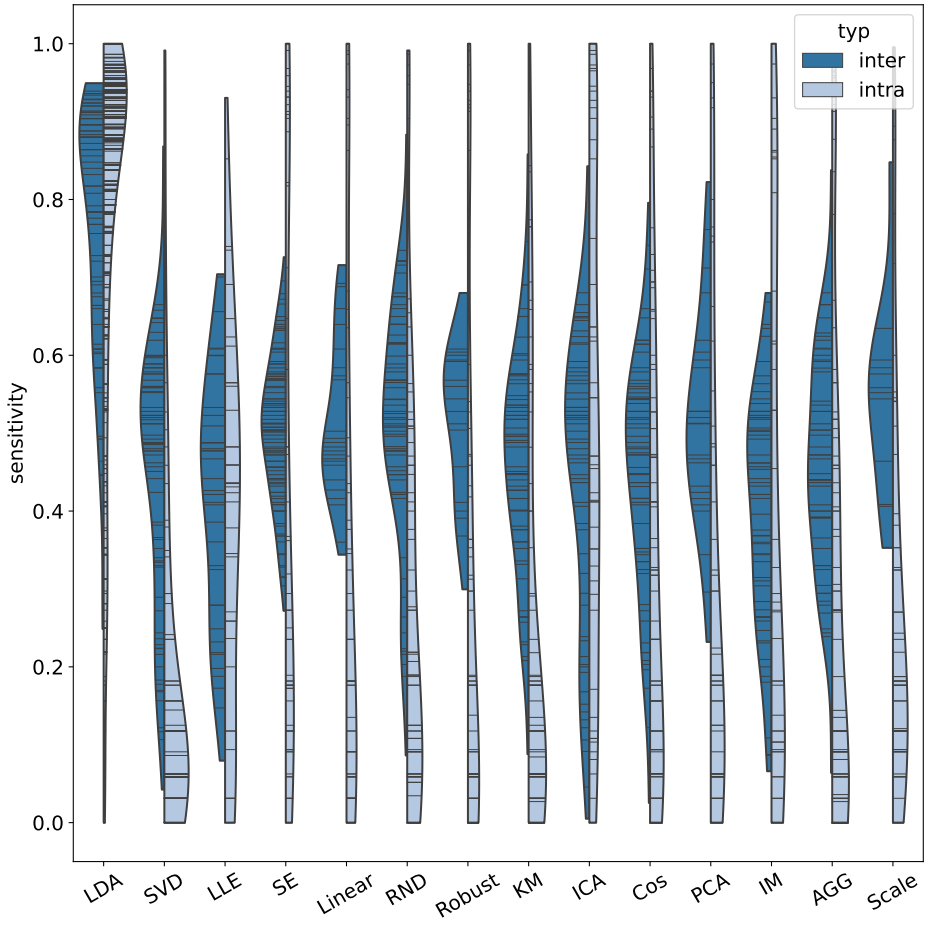

Figure 6.17.: Distribution of sensitivities associated with high accuracies of inter- and intra-participant analysis split by associated transformation. Intra-participant data obtained as previously discussed and unionized for display.

The sensitivity and specificity contributions of different transformations on this data, which are depicted in Fig. 6.17 and Fig. 6.18 respectively, yield extremely interesting results. Starting with the sensitivity, we find two particularly striking observations: Firstly, the peak of the sensitivity estimates for the respective configurations in the inter-participant dataset averages out around 50%, with a varying degree of variance for all transformations except LDA, which

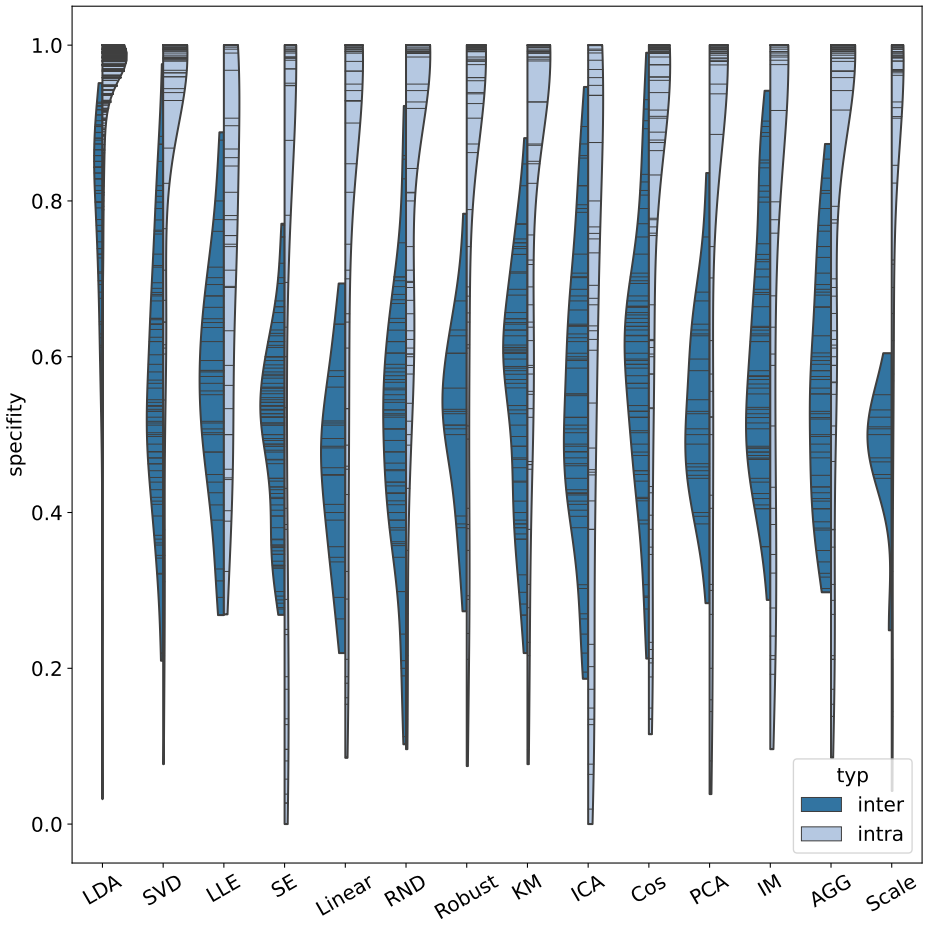

Figure 6.18.: Distribution of specificities associated with high accuracies of inter- and intra-participant analysis split by associated transformation. Intra-participant data obtained as previously discussed and unionized for display.

we will discuss separately in a moment. These variance measurement do not seem to be linked to the structural kind of transformation at play, with both the IsoMap or LLE as well as SVD or ICD yielding about comparable variances. The trend is visible for other transformations, including Spectral Embeddings, which have been included in the optimization results with high prevalence. Most to all transformations still improve their overall sensitivity estimate based on the configurations included for analysis compared to the intra-participant case, which do, however, perform substantially worse, trending towards very low sensitivities. The few exceptions of this are for one, LLE as well as k-Means, which show a significantly wider distribution. This is noticeable for k-Means in particular, which exhibits sensitivities at the bottom of the sensitivity scale, yet still manages to also place configurations at the top of it, whereas LLE only manages to peak at around 50% sensitivity, with just very few configurations including it as a data transformation exceed 60%. Furthermore IsoMap and Spectral Embeddings manage to place some configurations at the top above 80%, but the density here is also thin, and the peak of that distribution is located below 30% sensitivity. These results are, however only supported by

very few samples, as we have already discussed the preference of the optimizer to pick the LDA transformation for the intra-participant datasets, and this shows another clear evidence why this likely the case, as we, secondly, find that the LDA achieves peak sensitivities above the 90% mark in the intra-participant analysis, supported by a substantial amount of configuration samples, as well as achieving a peak of the sensitivity distribution even in the inter-participant case, only degrading the overall sensitivity distribution by about 5 percentage points. This is substantially less than the degradation in sensitivity when split by algorithm, again stressing the effectiveness of LDA as a helpful tool in improving classification performance not only with respect to accuracy, but also with a focus on recalling dysfunctional states from the dataset. Comparing this to the specificity analysis, we find that distributions do neither differentiate significantly in the intra-participant analysis, where all transformations are associated with very high specificities, nor in the inter-participant analysis, where peak specificities reach up to 70% to 90% specificity, but the distribution peak area remains solidly at about 50% to 70% specificity for all transformations, exhibiting substantial degradation in specificity compared to the intra-participant analysis. The only exception again is the LDA, which manages to exhibit a distribution that resembles the union of intra-participant results substantially better, only regressing to about 85% distribution peak specificity, showing once more its resilience for this kind of application and indicating again why it has been a preferable choice for the optimization procedure in the intra-participant optimization, even though it was not as prevalent when working on the inter-participant dataset.

6.2.4. Deconstructing accuracy emergence by classification scheme

Breaking down this sensitivity analysis over algorithms instead of transformations (again comparing against the union of equivalent results obtained during the intra-participant analyses), as it is depicted in Fig. 6.19, we see a split between two groups of algorithms: Firstly, we have the Gaussian Process as well as the Naive Bayes, which behave somewhat similar to the intra-participant case, covering almost the entire spectrum of possible sensitivities, with a peak of configurations actually below a sensitivity of 40%, however also a substantial fraction of configurations above 60%. This is a strong indication that for these two kinds of algorithms sensitivity does not contribute significantly to the achieved accuracies, hence, a significant contribution of to accuracy also originates in these algorithms' specificity, which is similarly depicted in 6.20 and shows a very similar behavior, the same two algorithms, Gaussian Processes and Naive Bayes, having a very widespread specificity distribution with a peak specificity around 80% and 60% respectively. This is approximately an inverted behavior compared to the sensitivity analysis, strongly indicating that specificity is overall slightly dominant with respect to the resulting optimal accuracy, showing that accuracy is overall slightly more driven by specificity than sensitivity in these, which reflect the findings for the intra-participant analysis, yet both algorithms are capable of achieving highest-level sensitivities. So for sensitivity-relevant applications both are still a viable option, but sensitivity must necessarily be included into the objective function for these kinds of applications. The other class of algorithms contains the tree-based schemes as well as the kNN. These exhibit substantially less variance both in the sensitivity distribution as well as the specificity distribution, coming out at around 50% for both, which is slightly less than the estimated accuracies of 60 to 70%, indicating that for these algorithms the accuracy behavior is actually supported equally by both the sensitivity as well as the specificity behavior. Hence, we find that these algorithms overall find a more balanced separation of the dataset in question that is for most configurations not significantly dominated by either, which also implies more robustness as different configurations do not sway the resulting sensitivities and specificities as significantly as it is the case for the Naive Bayes and the Gaussian Process schemes, which is an overall beneficial property to have when it comes to more widespread use. Within this group, we find the Decision tree to perform slightly better as the rest in terms of sensitivity, whereas

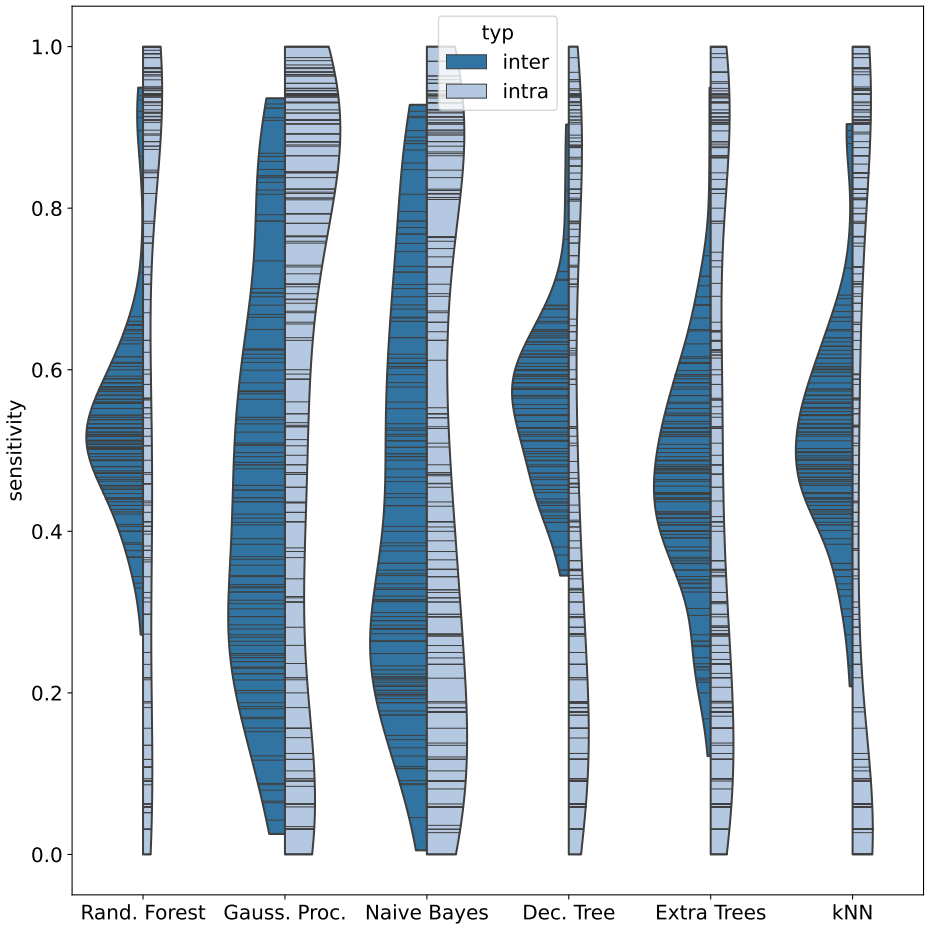

Figure 6.19.: Distribution of sensitivities associated with high accuracies of inter- and intra-participant analysis split by algorithm. Intra-participant data obtained as previously discussed and unionized for display.

the Extra-Trees have a heads up of about 20% in terms of specificity. All configurations exhibit about 90% specificity, but this seems to be the exception rather then the rule for this particular dataset and is hence less reliable for an even more diversified or differently created dataset.

Besides investigating the sensitivity distribution on the inter-participant dataset itself, we can also compare these results with the union of results that were obtained in the participant-specific analyses, both of which are already included in Fig. 6.19 as well as Fig. 6.20. This paints an interesting picture of both expected as well as unexpected results: The expected results include the degradation in overall specificity, an effect that pervades all algorithms to different degrees, as discussed already, where most algorithms are capable of reaching substantial specificities of solidly above 80%, most peaking at about 90%. This however, is not the case for the sensitivity comparison, where for the union of intra-participant datasets the accuracy is much less of a solid predictor, being rather widespread over the scale. In the inter-participant dataset, on the other hand, we find the two groups of algorithms, one of which exhibiting a very similar behavior, the other one exhibiting a much more compactified distribution that is less bimodal and closer to

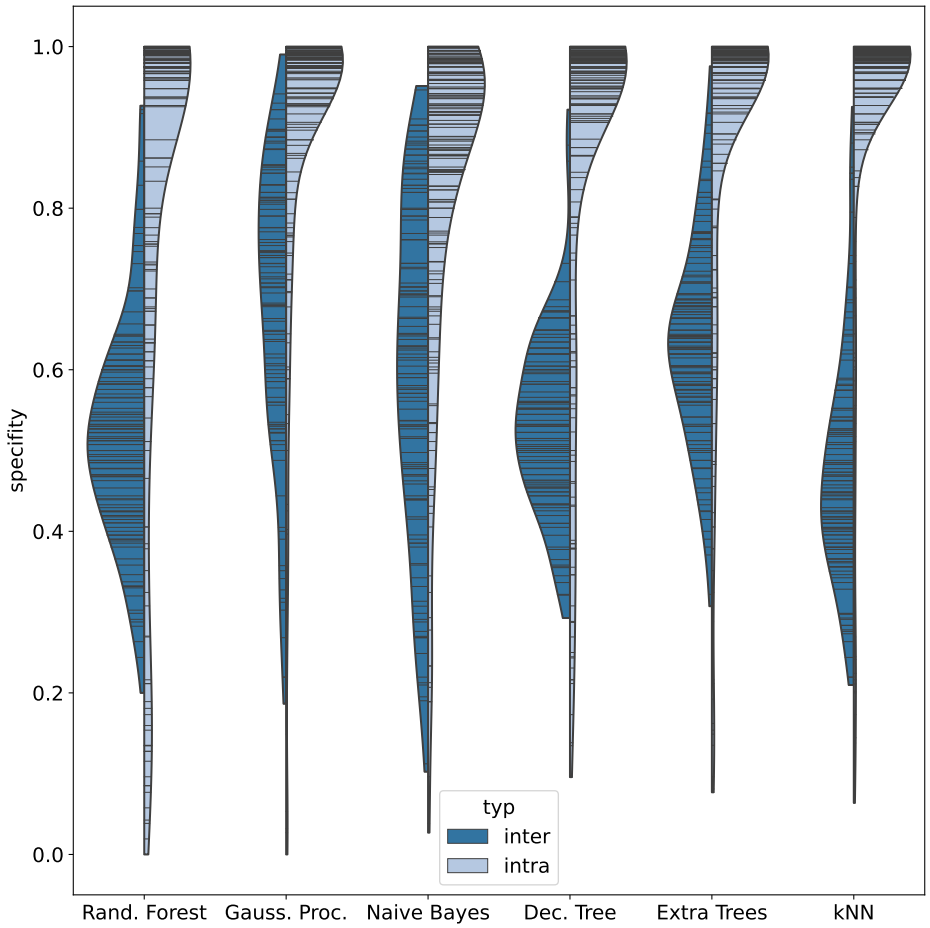

Figure 6.20.: Distribution of specificities associated with high accuracies of inter- and intra-participant analysis split by algorithm. Intra-participant data obtained as previously discussed and unionized for display.

the actual accuracy. This could be an early sign that despite the participant-wise individualism we have seen in the data this might not be an ultimate necessity, but instead we will see a trend towards overall convergence with a greater diversification of participant data. This is, however, a preliminary observation that needs further verification with more data than we currently have at our disposal.

Lastly, we will briefly review the effectiveness of different data fields on accuracy, depicted in Fig. 6.21. All data fields exhibit associations with accuracies that show about the same behavior we have discussed when investigating accuracies split by algorithm: We find a degradation in accuracy, which is unsurprising given the more diverse nature of the underlying dataset, but we also do not find substantial differences in terms of datafield effectiveness. The differences are even less pronounced compared to the algorithm-specific investigation. We again find that Fourier-transforms of the datafields do not yield any significant improvement in accuracy, which is consistent with what we could already observe in the intra-participant analysis. We have also investigated this question with regards to sensitivity, specificity and precision, which we did not

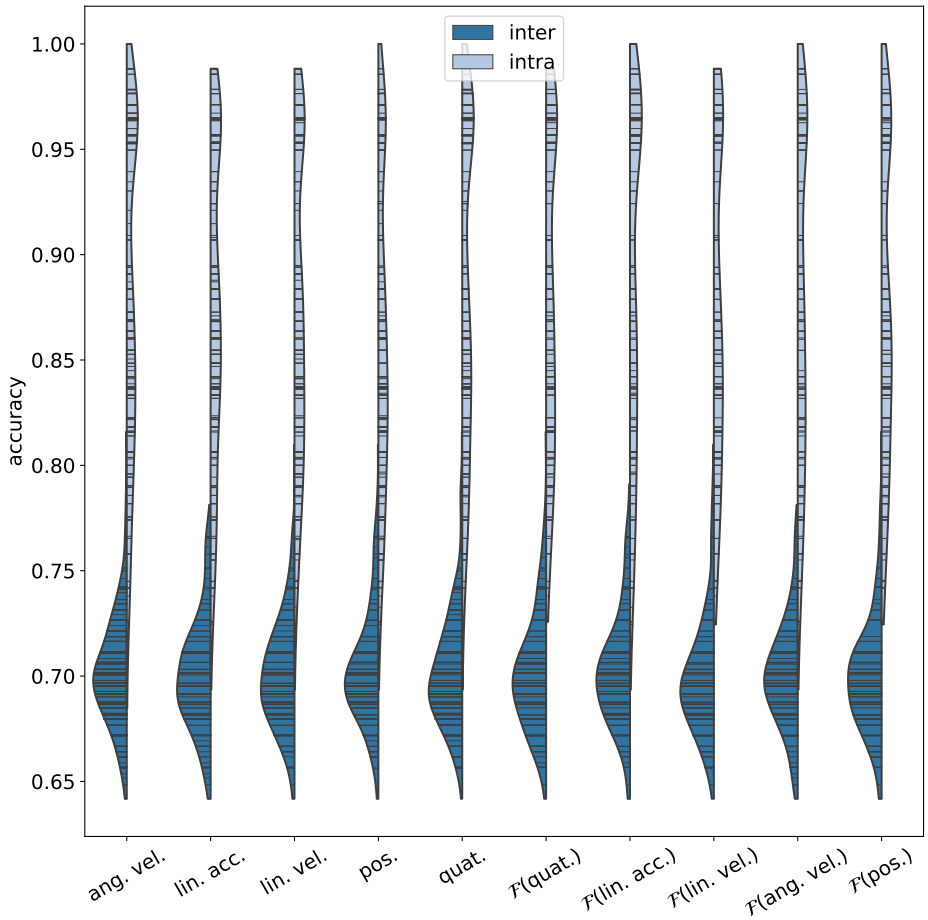

Figure 6.21.: Accuracy distribution of associated with data types, including Fourier-data based on them for both inter- and intra-participant analysis. Intra-participant analysis obtained by unionizing results from said analysis previously.

separately depict as the distributions of the different datafields are, just as the accuracy distributions, extremely similar, not indicating any preference whatshowever for any of the datafields. While this may seem as an unspectacular result, it does carry an important information for proceeding: As we do not find any improvement in classification accuracy or other properties by processing the base data the sensors yield, we can omit this processing step, working directly on the fundamental data, linear acceleration and angular velocity, simplifying the resulting procedure and reducing computational cost. This is particularly advantageous considering that the ultimate target platform for such a setup would be a totally integrated wearable device which is usually limited in computational power.

6.3. The impact of feature spaces for the classification problem

The analysis of the feature space representation component of the proposed procedure has yielded a number of key insights: We demonstrated that the proposed approach is actually capable of

yielding better-than-random classification accuracy with varying degrees of accuracy depending on the exact scenario and dataset.

We have investigated the relevance of classification schemes for the identification of dysfunctional cognitive states in participants diagnosed with major depressive disorder and identified several classification schemes as particularly effective in this. Among these are predominantly Gaussian Processes, which are algorithmically more complex, as well as the more simple k-Nearest-Neighbours. Several tree-based approaches have proven to be potentially interesting in future research as well. However, we have shown that the Gaussian Process and k-Nearest-Neighbour schemes yielded more consistent sensitivities, making them more attractive for the identification of dysfunctional cognitive states in patients diagnosed with major depressive disorder and proved to be more consistent in terms of precision as well, hence having a higher tendency to avoid misclassifications of such states by identifying dysfunctional states more correctly instead of just more frequent.

We also investigated the impact of base input data and have not found a substantial advantage of processed, Fourier-transformed or integrated data input fields over the data that is naturally provided by common IMU hardware, linear acceleration and angular velocity in both the individual studies as well as the cross-participant study. This means such processing can, for the application in major depressive disorder at least, be omitted, lightening the processing load on the overall system.

When investigating the effect of feature space transformations we have found the Linear Discriminant Analysis to repeatedly perform above the rest of the options in some form. While it was not disproportionally associated with higher accuracies for the cross-participant study compared to other feature space transformations, it yielded both the best accuracies in the individual investigation as well as consistently the best sensitivity values in all investigations, proving it to be highly beneficial particularly to the extraction of dysfunctional cognitive states. We could, also, clearly show that an additional processing on the data features is beneficial for the classification procedure, hinting at inherent structures in the feature space that can be pronounced for easier detection, which is in line with the performance of the Linear Discriminant Analysis that is specifically tailored to achieve this. Furthermore, certain details about this structure could already be inferred by the performance of structure-specific transformations that are tailored towards revealing hidden structures that are not well represented by the coordinate axis. Their not particularly exceptional performance combined with the noticeably exceptional performance of the Linear Discriminant Analysis can be a first indication that the structure we are seeking is in fact not hidden, but fairly well represented by the currently used coordinate system representation.

Overall, this chapter gave a strong indication that a well-suited preparation of the feature space can have a substantial impact on classification performance while at the same time allowing understanding and insight into how this performance comes into play and consequentially how the procedure can be simplified or accelerated by omitting non-useful components of the analysis in future work.

7. Classification configurations conceptually and practically

To conclude this part, we will do two things First, we will make a recommendation of classification to pick. In this context we will also investigate the hyperparameter distribution for the recommended algorithm. Second, we will summarize the general insights the results provide on a more abstract level.

7.1. An optimized classification scheme

Recalling the results so far, we can already make a substantial set of conclusions, some of them as a direct consequences of the results we discussed, others as a consequence of externalities. Firstly, in the intra-participant case we have seen a substantial contribution of the Linear Discriminant Analysis towards both good as well as stable classification. The LDA is not only prevalently picked by the optimization scheme, it is also significantly associated with high accuracies, making it a highly attractive choice for a classification scheme that not only facilitates high accuracies, but does so under particularly high sensitivities that are also robust under applied pretraining with a more general dataset. Furthermore, we have found that Fourier-transforms do not yield any noticeable improvement in classification accuracy, hence omitting them reduces the computational footprint of such an optimal classification scheme. Also, we have shown that different datafields do not imply significant improvements in classification, neither in the intra-participant nor in the inter-participant analysis, which allows us to reduce the computational footprint even further by omitting all datafields that require additional processing, such as velocities, or quaternions, and instead work with the data that is yielded by our sensor systems natively: accelerations and angular velocities.

With regards to classifiers, we found the two most reliable schemes to be Gaussian Processes and k-Nearest-Neighbours, each of which excelled in different aspects in detail and were found to be overall reliable and solid choices.

For the time being, we hence conclude that a k-Nearest-Neighbours classifier is the best choice for the problem of detecting dysfunctional mental states in patients diagnosed with depressive disorder due to two substantial advantages kNN has over the Gaussian Processes: Firstly, Gaussian Processes are known to be computationally more expensive than kNN, which is particularly relevant when it comes to running on embedded hardware. Also, kNN allows enhancements such as confidence regions and downsampling or resampling, which gives an easy way of controlling and further reducing the computational expense while retaining achieved accuracy, which in turn allows simple adaptation to the performance characteristics of the embedded hardware chosen. Furthermore, the kNN is the conceptually more simple algorithm, which is advantageous when it comes to the explainability of predictions made by it as both required by medical product regulations and advantageous for patient trust into the system. Tree-based algorithms have found to be of noticeable effectiveness, but lack the simplicity of a k-Nearest-Neighbours approach and are at the same time almost always outperformed by a Gaussian Process approach.

We can therefore conclude, that a

- k-Nearest-Neighbours with

- angular velocity and acceleration data
- enhanced by a Linear Discriminant Transformation

seems like the most advantageous construct for the goals of the objective we set out to solve. Finally, we need to take a look at hyperparameters that were also part of our proposed optimization procedure. We will limit ourselves to discussing this for our proposed classification scheme, the k-Nearest-Neighbours, specifically and exemplarily.

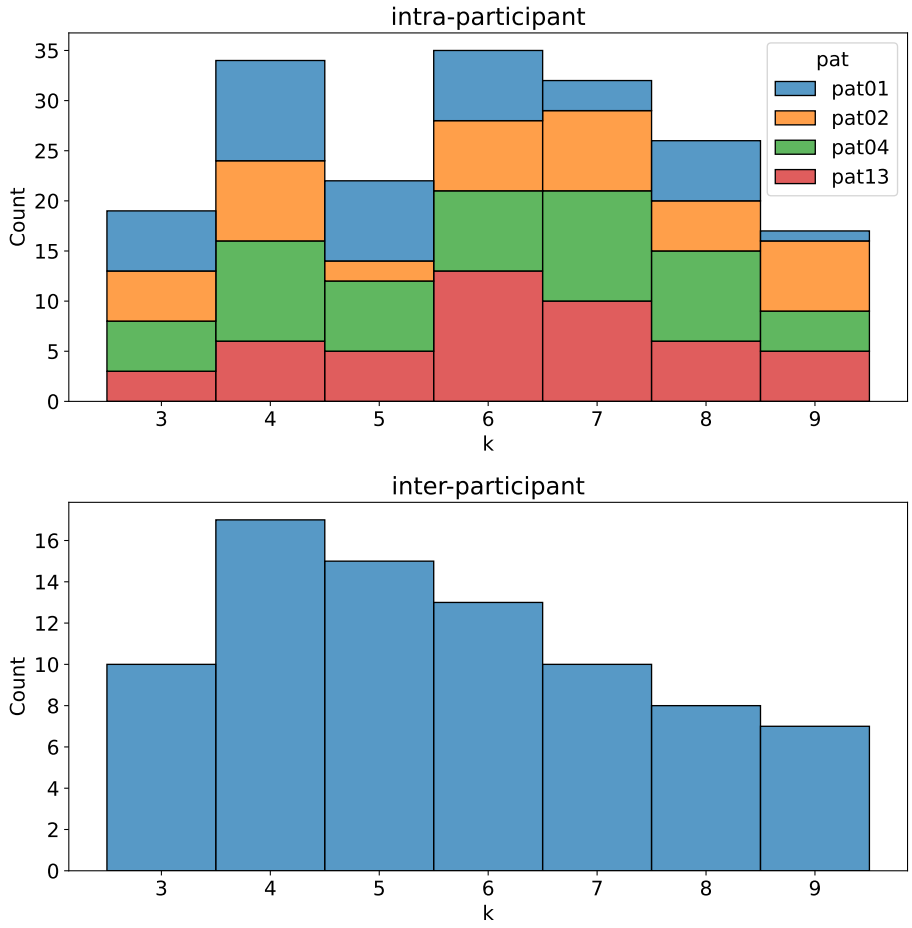

Figure 7.1.: Prevalence of hyperparameter values for the number k of neighbours for the k-Nearest-Neighbours classification scheme

The k-Nearest-Neighbours in our parametrization has exactly one hyperparameter, the k number of neighbours, the distribution of which we found among the commonly discussed set of configurations in this work is depicted in Fig. 7.1 for all optima of successive optimization runs. The space of possible k during the optimization was set such that

$$k \in [3, \cdots, 9] \subset \mathbb{N} \, . \tag{7.1}$$

The lower cap was set to ensure two things:

1. reduce the level of potential overfitting by requiring not just one sample reference close-by, but multiple references

2. avoid a possible tie at the boundaries, which for a two-label-problem could happen for $k = 2$, but not for $k = 3$

The upper cap equivalently was set to also adhere to the second condition while also constraining our search space for efficiency purposes. We do find a distribution of k that for both the inter- as well as the intra-participant dataset is falling towards the sides, which is a strong indicator that we are indeed within the convergence area. Furthermore, we see that, while the intra-participant dataset shows a slightly wider distribution core, ranging from $k = 4$ to $k = 8$, the inter-participant data shows a distribution mode at $k = 4$ with successive k monotonically declining, with $k = 7$ showing approximately the same prevalence as $k = 3$. This strongly implies that, for one, k should not be selected too large to still being able to capture less populated regions in feature space and consequentially retain good classification performance, but also not too small to account for the inevitable effect of overfitting at too small k. Inspecting the two distributions, $k = 5$ seems to be an excellent choice overall, bringing both aspects together while at the same time ensuring to always be capable of breaking at tie for samples close to classification area borders for the two-class-problem at hand.

We finally find that this way we can achieve accuracies that rely on native IMU data of up to 93%, 89.4%, 97.8% and 80% for the pat01, pat02, pat04 and pat13 datasets respectively based solely on motion data leveraging k-Nearest-Neighbours scheme.

7.2. Generalized insights into optimizing classification configurations

In this part, we have approached the automated derivation of classification procedures that are capable to extract implicit characteristics from IMU measurement data to differentiate between the acute presence of distinct, highly aversive symptoms of depression and the absence thereof, and also to allow an understanding of how this differentiation was achieved. We have applied this approach to quantitatively detect indicators for dysfunctional cognitive states in the context of major depressive disorder, which to our knowledge has never been attempted in the context of major depressive disorder yet.

To do so we have separated a feature extraction procedure from the classification procedure, allowing us to assess both components separately. First, the feature extraction process was divided into data slicing, property derivation from said data slices and a transformation of the resulting property vector to increase separability. In a second, separate step we have applied classical Machine Learning techniques to the resulting feature space.

We have performed this procedure both on non-mixed datasets as well as on a dataset containing multiple different participants. We find both a degradation in accuracy when the data is no longer related to a single individual as well as a still higher than random identification of depressive states. This is a strong indicator that depressive symptoms occur with substantial individual variation, but that there still is an underlying structure that can be generalized over participants. This is consistent with the diagnostic procedures for major depressive disorder, where different participants exhibit different behavior leading to the same diagnosis, but the list of potential behaviors is still finite and hence acknowledges the existence of underlying structures.

We could further show that our procedure not only allows for automated derivation of an optimized classification procedure, but that the separation of components provided insight into the structural systematic of the underlying data, even on the lower end of the amount of data

used. We could identify that for the case of depressive disorder, the non-transformed IMU sensor data is already well-suited for differentiating between the presence of rumination as a distinct depressive symptom and its absence and have found that simple representations thereof already contribute greatly to this differentiation.

Additionally, our method not only allowed for the autonomous derivation of an optimized classification procedure, but also for the identification and contribution of different components to this classification result. This is particularly relevant when considering the sensitivity and specificity of different components, e.g. the contribution of correct identifications of acute depressive symptoms in contrast to the predominantly correct identification of their absence. Particularly the Linear Discriminant Analysis has been shown to overall allow for extremely high sensitivities and specificities both when used on individual datasets as well as when used on inter-participant datasets. As this can make a significant difference for the practical application, depending on the way it is integrated into therapeutic measures, we propose to not use a standard objective for this optimization, but to tailor the objective specifically to the needs and requirements of the desired application.

Part III.

Towards optimal measurement of motion

So far, type, number and location of sensors and their acquired data was assumed to be predetermined However, this existing data might also be obtained in more or less advantageous configurations, contributing to or ablating from better separability of classes or better insight into the underlying mechanics that determine human motion. Also, we have found during the study, in which the data that formed the base for the analysis in the previous part was obtained, that a more complex measurement apparatus reduces participant acceptance, limiting the amount of data that can be obtained in the first place. Hence, being able to acquire the same amount of information with a less obtrusive system is highly desirable both for participant acceptance, but also for ease of operation and cost of deployment, for example being able to deploy fewer sensors that overall capture the same information by intelligently locating these fewer sensors. However, less sensors are not always an improvement, as the removal of a sensor capturing a crucial part of the information is likely to lead to a substantial decrease in achieved performance, whichever indicator for this performance is chosen.

Therefore, we will now investigate the impact of different sensors on a motion reconstruction problem, as well as deriving methodology to optimize the location of a given sensor set. The foundational problem we want the sensor set to be optimized for can ultimately be chosen fairly arbitrarily, such as a classification problem originating in the previous investigation, given it is sufficiently parameterized. As this particular problem already comes with a substantial incumbent complexity, we will instead derive the approach for a slightly simpler model-driven motion reconstruction problem. The objective of this particular problem is the reconstruction of a given motion, of which IMU motion data was recorded, hence working with similar data inputs as were used in the previous investigation.

We will assess the importance of individual sensors in this arrangement and their effect on the degrees of freedom of the underlying model as well as deriving a procedure to establish an optimized sensor layout for this reconstruction to work, both under ideal circumstances as well as under the impact of non-perfect sensors.

We will also show that this approach can be used to determine the suitability of sensors of different quality and typically therefore price for a defined test problem. This can be used to both attest to the suitability of an available sensor for a defined objective problem as well as allowing to establish bounds on the required parameters if a choice of sensor has not been made, thinning out the field of candidates and allowing to establish a Pareto-optimum, for example by availability or price and quality of solution of the objective problem.

This can then be used in combination with an approach like in the last part, to determine a set of sensors that both yield sufficient amounts of data to allow for the identification of aversive cognitive states while at the same time retain patient compliance and increasing acquisition time with regards to battery life by avoiding unnecessary sensors. This can be done both in terms of assessing the relevance of existing sensors, or, with sufficient data to model a continuous surface, also derive an optimal layout and locations for the sensors in use.

8. Model-driven reconstruction of motion

Recording a motion can be done in various ways: The most well-known one is marker-based motion recording, which is commonly used in the creation of movies and computer games to map human-like motion to computer-animated character models. For this, a participant is equipped with reflective markers on specific points of the body, which are then illuminated and recorded by an array of cameras surrounding that participant. If the position of the cameras relative to each other are known, this allows the localization of every marker in the volume covered by the cameras. These position information can then be used to reconstruct the motion afterwards, either solely by evaluating marker trajectories or by projecting those marker trajectories onto a model of the body of the participant. By fitting the trajectories of the markers to the recorded trajectories returns the according motion of the model. While this approach is very popular due to its high precision, it comes with a number of disadvantages, the most prominent one being the confinement of the recording area to what is well covered by the according cameras, which severely limits its usecase outside of a prepared laboratory environment.

In contrast to this, IMU-based systems measure local acceleration and angular velocity at points where these IMUs are strapped onto the participants body. The principle is fundamentally the same, measuring acceleration and angular velocity curves and fitting a model to those. The model-based reconstruction is noticeably more important in this case, however, as IMU-based motion reconstruction is less precise compared to optical markers for a number of reasons: Firstly, whereas cameras are static and can therefore be calibrated with high precision and the according markers are typically located at specific bone landmarks and rotationally invariant, this is noticeably more difficult for IMUs, being not rotationally invariant and commonly not small enough to be precisely located on specific bone landmarks. Also, as IMUs are the primary measurement device in contrast to marker-based systems, where the measurement happens on the cameras, every measurement is subject to both measurement noise as well as a monotonically increasing accumulation of offset-errors in integrated data properties such as positional information or orientation, increasing the error on the reconstructed positions and consequentially the entire trajectory over time.

Proof. Be $\ddot{x}(t)$ the acceleration of the located IMU. Its trajectory $x(t)$ can then be computed as

$$x(t) = \int_0^t \int_0^{t'} \ddot{x}(t'') \, \mathrm{d}t'' \, \mathrm{d}t' \ . \tag{8.1}$$

However, in reality the measurement of $\ddot{x}(t)$ is noisy, with a systematic error μ and a statistical error σ, for which we know

$$\int_{-\infty}^{\infty} \mu > 0 \tag{8.2}$$

$$\int_{-\infty}^{\infty} \sigma = 0 \ . \tag{8.3}$$

This has consequences for eq. (8.1), as we need to rewrite our measured acceleration as

$$\ddot{x}_m(t) = \ddot{x}(t) + \mu + \sigma \tag{8.4}$$

and consequentially, eq. (8.1) changes to

$$\begin{aligned}
x(t) &= \int_0^t \int_0^{t'} \ddot{x}_m(t'') \, \mathrm{d}t'' \, \mathrm{d}t' && (8.5)\\
&= \int_0^t \left(\int_0^{t'} \ddot{x}(t'') + \mu + \sigma \, \mathrm{d}t'' \right) \mathrm{d}t' && (8.6)\\
&= \int_0^t \left(\int_0^{t'} \ddot{x}(t'') \, \mathrm{d}t'' + \int_0^{t'} \mu \, \mathrm{d}t'' + \int_0^{t'} \sigma \, \mathrm{d}t'' \right) \mathrm{d}t' && (8.7)\\
&= \int_0^t \left(\int_0^{t'} \ddot{x}(t'') \, \mathrm{d}t'' + C + 0 \right) \mathrm{d}t' && (8.8)\\
&= \int_0^t \int_0^{t'} \ddot{x}(t'') \, \mathrm{d}t'' \, \mathrm{d}t' + Ct \ . && (8.9)
\end{aligned}$$

We see that while the statistical noise is less of a problem, the systematic error on the position, which can for example originate in manufacturing tolerances with in the IMU, increases monotonically with time. □

This is commonly referred to as drift and in contrast to the marker-based systems, where this kind of error is not integrated up as positional information is measured directly and hence not amplifying over time, it contributes to an increasing error on the positions over time. The same error on marker-based positions remains constant, i.e.

$$x_m(t) = \ddot{x}(t) + \mu + \sigma \ . \tag{8.10}$$

The most effective way of tackling this is constraining the involved IMUs with an underlying model that requires a certain physiology. This way, IMUs drifting into different directions can be used to correct each other. Depending on the exact correction strategy, only the underlying average drift over all sensors remains.

8.0.1. Rigid body modelling of kinematic trees

We will use a model-based approach to reconstruct a predefined motion from IMU sensor data. The underlying model is a rigid multi-body model of the upper body, which is depicted in Figure 8.1 and set up with the specific measurements of the participant performing the motion for better reconstructive accuracy [24], consisting of ten individual rigid bodies which are connected via joints, constituting a kinematic tree. The motion we intend to reconstruct is depicted in Figure 8.2. We will follow the parametrization of [22] to describe each individual rigid body in 6D coordinates, unifying linear and angular components into one vector. Specifically, we write the 6D-velocity of a body with linear velocity $\boldsymbol{v}$ and angular velocity $\boldsymbol{\omega}$ in the form

$$\hat{\boldsymbol{v}} = (\boldsymbol{\omega}, \boldsymbol{v})^T \ , \tag{8.11}$$

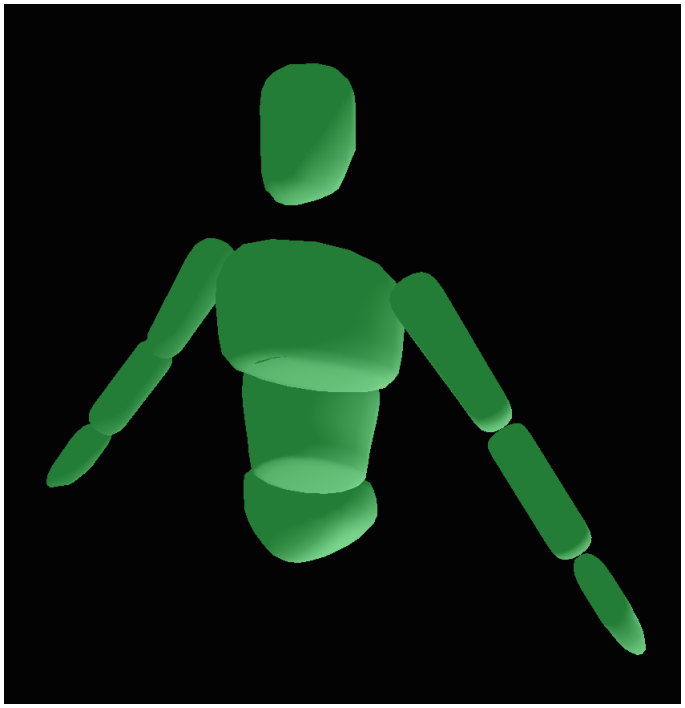

Figure 8.1.: The upper-body segmented model.

which allows for more efficient computation of both properties in one step. To achieve this, we need to rewrite the traditional rotational matrix $R \in \mathbb{R}^{3\times3}$ and translation vectors $r \in \mathbb{R}^3$ for a body into a combined transformation T, which reads

$$T(R,r) = \begin{pmatrix} R & 0 \\ -Rr\times & R \end{pmatrix} \in \mathbb{R}^{6\times6} \ . \tag{8.12}$$

This allows to transform a motion vector $\hat{\boldsymbol{v}}$ from coordinate frames A to B via

$$\hat{\boldsymbol{v}}_B = T(R_B, r_b)\hat{\boldsymbol{v}}_A \tag{8.13}$$

This now allows the construction of a kinematic tree where bodies are interconnected. The body representing the left hand, B_1, is connected at a defined point to the body representing the left forearm, B_2, which is in turn connected to the body representing the left upper arm, B_3. If B_3 is moved, this consequentially affects the bodies down the kinematic tree, as they are translated or rotated accordingly. Therefore, we need to express the transformations for each dependent body based on its parent body to find the transformation $\hat{T}$ into the body-local coordinate system from the global coordinate system, which for bodies B_1 and B_2 read, from the point of B_3 as

$$\hat{T}_2 = T_2\hat{T}_3 \tag{8.14}$$

$$\hat{T}_1 = T_1T_2\hat{T}_3 \tag{8.15}$$

or more generally

$$\hat{T}_i = \prod_{j=0}^{j=i} T_j \ . \tag{8.16}$$

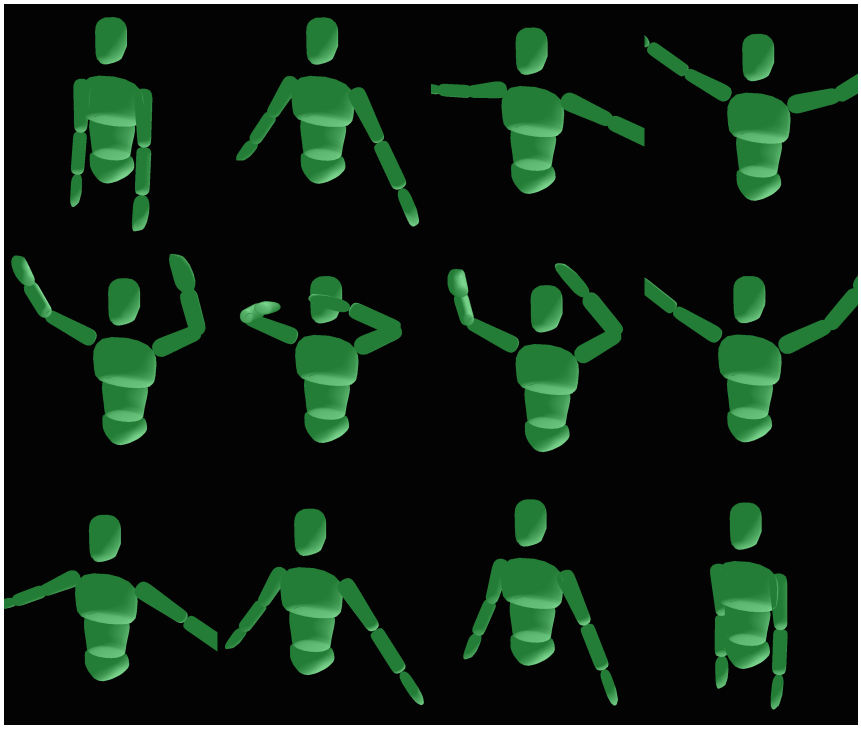

Figure 8.2.: The sequenced motion used as a basis to develop the following procedure.

To restrict the motions possible on each joint of the model, we follow [23] to additionally constrain relative motion between two bodies by introducing motion subspace matrices $S \in \mathbb{R}^{6\times n}$ for the joints, with $n \in [1, \ldots, 6]$ being the degrees of freedom a particular joint allows, which specify which parts of the motion vector can be accentuated independently from the parent body and which cannot. A joint that allows free relative rotational motion in ω_x would therefore read

$$S_{\omega_x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} , \tag{8.17}$$

whereas a freely rotating joint would have a corresponding

$$S_{\omega_x,\omega_y,\omega_z} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} . \tag{8.18}$$

This allows for a set of accentuations $\boldsymbol{q} \in \mathbb{R}^n$ to automatically map to the accentuated dimensions

base body name	rot. d.o.f.	transl. d.o.f.
Pelvis	x, y, z	x, y, z
L5_L3	x, y, z	none
T12_T8_Shoulders	y, z	none
Head_Neck	y	none
RightUpperArm	x, y, z	none
RightForeArm	y, z	none
RightHand	x	none
LeftUpperArm	x, y, z	none
LeftForeArm	y, z	none
LeftHand	x	none

Table 8.1.: List of bodies and their rotational and translational degrees of freedom.

of a joint and allows expressing joint velocities and accelerations as

$$\hat{\boldsymbol{v}}_j = S_j \dot{\boldsymbol{q}}_j \tag{8.19}$$

$$\hat{\boldsymbol{a}}_j = \frac{d}{dt}\hat{\boldsymbol{v}}_j = \dot{S}_j \dot{\boldsymbol{q}}_j + S_j \ddot{\boldsymbol{q}}_j \ . \tag{8.20}$$

For a detailed derivation of analytical expressions of these subspace matrices, we refer to [23].

Based on the set of accentuated joints, the underlying rigid body model is now capable of representing motion by specifying a vector $\boldsymbol{q}(t) \in R^{n_M}$ of according length in both time and number of total degrees of freedom n_M of the model. We will apply this to a ten-segment rigid multibody model as depicted in Fig. 8.1 with 21 rotational degrees of freedom, which are listed in Tab. 8.1.

The motion depicted in Fig. 8.2 was recorded with the Xsens sensor system [86] in form of a full-body motion and transferred to the aforementioned model and the trajectory of the corresponding degrees of freedom were subjected to a denoising procedure of a two-way butterworth filter to yield a maximally clean, yet human-like motion.

8.0.2. Integrating IMU sensors into a rigid body model

To use this model for motion reconstruction, we also need to parameterize IMUs on that model. Generally, the position of an IMU is defined by the base body it is applied to, as well as a position vector $\vec{p} \in \mathbb{R}^3$ from a defined point of that base body, for example the origin. We will for the moment not model relative rotations of IMU to its base body. Furthermore we will reduce the dimensionality of the problem by not requiring

$$\vec{p} \in \mathbb{R}^3 \tag{8.21}$$

but instead incorporate the constraint of locating the IMU noninvasively on the surface of the segment to achieve

$$\vec{p} \in \mathbb{R}^2 \tag{8.22}$$

by reparametrizing as follows: We first approximate the surface of rigid bodies that make up the kinematic tree model as conical frustrums. By knowing the dimensions on the underlying segments, we can define a width in both x and y direction, both at the top as well as at the bottom of the segment, as some segments might taper. While the approximation as conical frustrums is certainly a simplification, it will allow the reduction of a volume parametrization to a surface

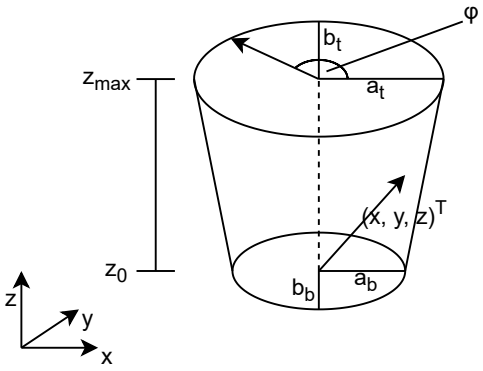

Figure 8.3.: The cone model.

parametrization, which we will assume to be sufficiently accurate for the moment. The conical frustrum model can later be easily replaced by a more sophisticated surface parametrization without changing its degrees of freedom. With the measurement labels depicted in Fig. 8.3, we can express any position that is described by a vector $(x, y, z)^T$ as a vector $(\phi, z)^T$ from the origin of the body to a point on the model surface via

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \left(\frac{a_t - a_b}{z_{max} - z_0}(z - z_{max}) + a_t\right) a_t \sin(\phi) \\ \left(\frac{a_t - a_b}{z_{max} - z_0}(z - z_{max}) + a_t\right) b_t \cos(\phi) \\ z \end{pmatrix} . \tag{8.23}$$

Proof. We will first describe an arbitrary point $(x, y, z)^T$ in generalized cylindrical coordinates as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \sin(\phi) \\ b \cos(\phi) \\ z \end{pmatrix} \tag{8.24}$$

for a and b describing appropriate width and depth for a horizontal slice through the cone at height z. We furthermore assume without loss of generality that

$$a_t > a_b \ . \tag{8.25}$$

We notice that by construction of the model as a cone frustrum, the flanks of the model are linearly descending from a_t to a_b, i.e. we can express this descent as a linear transition for any point $x_i(z_i)$ as

$$x_i(z_i) = c_1 z_i + c_2 \tag{8.26}$$

and furthermore, we know that

$$\frac{x_i(z_i)}{x_t} = \frac{a_i(z_i)}{a_t} \tag{8.27}$$

$$\iff x_i(z_i) = \frac{x_t}{a_t} a_i(z_i) \tag{8.28}$$

by construction of the conic nature of the frustrum, with $a_i(z_i)$ describing the width of the model at height z_i. We know that by construction of our polar description

$$x_t = a_t \sin(\phi) \tag{8.29}$$

and that for a we find

$$a_t = c_1 z_{max} + c_2 \tag{8.30}$$
$$a_b = c_1 z_0 + c_2 \tag{8.31}$$
$$\tag{8.32}$$

which by simple reformulation leads to

$$c_1 = \frac{a_t - a_b}{z_{max} - z_0} \tag{8.33}$$
$$c_2 = a_t - c_1 z_{max} = a_t - \frac{a_t - a_b}{z_{max} - z_0} z_{max} \tag{8.34}$$

Using this for the general expression of $a(z)$ yields

$$a(z) = \frac{a_t - a_b}{z_{max} - z_0}(z - z_{max}) + a_t \tag{8.35}$$

and consequentially, combining eq. (8.35) with eq. (8.28) we find

$$x(z) = \frac{x_t}{a_t} \frac{a_t - a_b}{z_{max} - z_0}(z - z_{max}) + a_t \ . \tag{8.36}$$

The derivation for $y(z)$ yields an according result, which, due to symmetry considerations, i.e.

$$\frac{a_t}{a_b} = \frac{b_t}{b_b} \ , \tag{8.37}$$

translates to the same rescaling factor, such that we can generally express

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \left(\frac{a_t - a_b}{z_{max} - z_0}(z - z_{max}) + a_t \right) a_t \sin(\phi) \\ \left(\frac{a_t - a_b}{z_{max} - z_0}(z - z_{max}) + a_t \right) b_t \cos(\phi) \\ z \end{pmatrix} \ , \tag{8.38}$$

with the z-component staying unchanged between both coordinate systems.

□

This expression now allows us to express an IMU setup in a space $\frac{2}{3}$ of its original size in terms of positional dimensions. We do not apply relative rotations of the IMU to the base body at this point, but will assume the IMU to be oriented along the frame of the base body it is mounted on.

8.1. Reconstruction as an optimal control problem

In the following we will derive the framework of optimal control problems, as well as an approach of solving them. While we could use a frame-by-frame fitting procedure to reconstruct motion at specific measurement points, this approach has the substantial disadvantage of not correlating neighbouring points with each other. As a consequence, this approach, while simple, yields the risk of discontinuities in the resulting solution, making the motion unviable. This is particularly risky with regards to overfitting, as individual frames might yield a better fit that is discontinuous with the neighbouring reconstructed frames particularly when noise is applied. Approaching such a reconstruction as an optimal control problem yields means to ensure the continuity of the resulting solution, guaranteeing a biomechanically viable motion in our particular application and a time-consistent, continuous solution in the more general, problem-unspecific case.

8.1.1. General formulation of optimal control problems

To set up a single-phase optimal control problem, we will first define state and control variables as well as static parameters of the problem

$$\boldsymbol{x}(t) \in \mathbb{R}^{n_x} \tag{8.39}$$
$$\boldsymbol{u}(t) \in \mathbb{R}^{n_u} \tag{8.40}$$
$$\boldsymbol{p} \in \mathbb{R}^{n_p} \ . \tag{8.41}$$

Based on this, we can define the underlying system dynamics as an ordinary differential equation for the state variables

$$\dot{\boldsymbol{x}} = f(t, \boldsymbol{x}(t), \boldsymbol{u}(t)) \tag{8.42}$$

which we define on a time interval $\mathcal{T} := [t_0, T] \subset \mathbb{R}$. We furthermore specify, most generally, a Bolza-type objective function that is to be minimized:

Definition 4. *We define an infinite-dimensional Bolza-type objective function as*

$$\Phi(\boldsymbol{x}, \boldsymbol{u}(t), \boldsymbol{p}(t)) = \int_{t_0}^{T} \Phi_L(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p})\, dt + \Phi_M(T, \boldsymbol{x}(t), \boldsymbol{p}) \tag{8.43}$$

consisting of a Lagrange-type objective function Φ_L, which emphasises the minimization of a property over the duration of the time interval in question, and a Mayer-type objective function Φ_M, which emphasises the minimization of a property at the end of the time interval in question.

Consequentially, this infinite-dimensional objective function results in an infinite-dimensional optimization problem. Both Lagrange- and Mayer-type formulations can typically be transformed into one another and are hence exchangeable, however, some problems are just formulated more naturally as one of the two types. In addition to this, we define four types of constraints:

Definition 5. *Path constraints*
We define path equality constraints $\boldsymbol{g}$ and path inequality constraints $\boldsymbol{h}$ such that

$$\boldsymbol{g}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}) = 0 \ \forall \ t \in \mathcal{T} \tag{8.44}$$
$$\boldsymbol{h}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}) \geq 0 \ \forall \ t \in \mathcal{T} \tag{8.45}$$

specify constraints that apply for each point over the entire time interval $\mathcal{T}$.

Definition 6. *Point constraints*
We define point-wise equality constraints r_{eq} and point-wise inequality constraints r_{ineq} that specify constraints at specific points in time as

$$\boldsymbol{r}_{eq}(\boldsymbol{x}(t_0), \ldots, \boldsymbol{x}(T), \boldsymbol{u}(t_0), \ldots, \boldsymbol{u}(T), \boldsymbol{p}) = 0 \tag{8.46}$$
$$\boldsymbol{r}_{ineq}(\boldsymbol{x}(t_0), \ldots, \boldsymbol{x}(T), \boldsymbol{u}(t_0), \ldots, \boldsymbol{u}(T), \boldsymbol{p}) \geq 0 \ . \tag{8.47}$$

This consequentially leads us to

Definition 7. *An infinite-dimensional constrained optimal control problem is defined as*

$$\min_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{p}} \quad \Phi(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p}) \tag{8.48a}$$

$$\text{subject to} \quad \dot{\boldsymbol{x}} = f(t,\boldsymbol{x}(t),\boldsymbol{u}(t)) \tag{8.48b}$$

$$\boldsymbol{g}(t,\boldsymbol{x}(t),\boldsymbol{u}(t),\boldsymbol{p}) = 0 \;\forall\; t \in \mathcal{T} \tag{8.48c}$$

$$\boldsymbol{h}(t,\boldsymbol{x}(t),\boldsymbol{u}(t),\boldsymbol{p}) \geq 0 \;\forall\; t \in \mathcal{T} \tag{8.48d}$$

$$\boldsymbol{r}_{eq}(\boldsymbol{x}(t_0),\ldots,\boldsymbol{x}(T),\boldsymbol{u}(t_0),\ldots,\boldsymbol{u}(T),\boldsymbol{p}) = 0 \tag{8.48e}$$

$$\boldsymbol{r}_{ineq}(\boldsymbol{x}(t_0),\ldots,\boldsymbol{x}(T),\boldsymbol{u}(t_0),\ldots,\boldsymbol{u}(T),\boldsymbol{p}) \geq 0 \tag{8.48f}$$

$$\tag{8.48g}$$

with $\Phi(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p})$ defining an objective function, $\boldsymbol{x}(t)$ defining a state variable vector, $\boldsymbol{u}(t)$ defining a control variable vector, $\boldsymbol{g}$ and $\boldsymbol{h}$ defining path equality and inequality constraints respectively and $\boldsymbol{r}_{eq}$ as well as $\boldsymbol{r}_{ineq}$ respectively describing point equality and inequality constraints.

8.1.2. The Direct Multiple Shooting Method

To solve the constrained infinite-dimensional optimal control problem defined in Def. 7, we will employ the Direct Multiple Shooting Method [7] via its implementation by [61] in the form of MUSCOD-II. As Direct Multiple Shooting is a discretize-then-optimize approach, in contrast to indirect optimize-then-discretize methods, we will first discretize the problem both in state and control variables. We will then derive the underlying initial value problems and the necessary constraints that ensure a smooth resulting solution to obtain the discretized finite-dimensional constraint optimal control problem.

Generally, the Direct Multiple Shooting Method segments the interval $[t_0, T]$ into a number of sub-intervals and discretizes both control and state parametrizations according to these intervals, creating an initial value problem (IVP) on all intervals. In addition to that, Direct Multiple Shooting employs continuity conditions on interval transitions, ensuring a smooth final trajectory. We will derive the individual components of this in the following.

Discretization of controls $\boldsymbol{u}(t)$

To discretize the controls $\boldsymbol{u}(t)$ we first define a *control grid*

$$G_u = (t_0, \cdots, t_j, \cdots, t_{n_u} = T) \subset \mathcal{T} \tag{8.49}$$

that consists of a sequence of time points which will constitute the start of a piecewise approximation of the control function

$$u_j(t) \approx u_j(t, \boldsymbol{c}_j) = \phi(t_j, \boldsymbol{c}_j) \tag{8.50}$$

that is defined on the interval $[t_j, t_{j+1}]$. These control functions can be chosen flexibly, but common choices are a piecewise constant approximation of $\boldsymbol{u}(t)$, i.e.

$$\phi(t) = c_j \text{ for } t_j \leq t \leq t_{j+1} \tag{8.51}$$

or piecewise linear, i.e.

$$\phi(t) = \frac{t - t_j}{t_{j+1} - t_j} c_{j,1} + c_{j,2} \text{ for } t_j \leq t \leq t_{j+1} \tag{8.52}$$

Alternatively, higher order base functions for the control approximation, such as higher order polynomials or splines, are also possible to be used with Direct Multiple Shooting. Whereas

piecewise constant approximations of the controls will inherently result in a discontinuous approximation (unless $c_i = c_j \forall i, j \in [0, \cdots, n_u]$), higher order approximations can be constrained to yield a continuous control approximation by additionally requiring

$$\phi_j(t_{j+1}) = \phi_{j+1}(t_{j+1}) \ . \tag{8.53}$$

Discretization of state parametrizations $\boldsymbol{x}(t)$

The next step is the discretization of the differential states $\boldsymbol{x}(t)$, for which we define a state grid analogously to the control grid before, i.e.

$$G_x = (t_0, \cdots, t_i, \cdots, t_{n_x} = T) \subset \mathcal{T} \tag{8.54}$$

on which we can leverage the specified state derivative $\dot{\boldsymbol{x}}$ to constitute a set of initial value problems (IVPs) that read

$$\dot{\boldsymbol{x}} = f(t, \boldsymbol{x}(t), \boldsymbol{u}(t)) \tag{8.55}$$

$$\boldsymbol{x}(t_i) = \boldsymbol{s}_i \tag{8.56}$$

on the interval $[t_i, t_{i+1}]$, with s_i denoting the initial value of the IVP at t_i. This IVP can then be solved with standard procedures to obtain the trajectory of $\boldsymbol{x}(t)$ for $t_i < t < t_{i+1}$ for each interval on the state grid, an exemplary result of which can be seen in Fig. 8.4 on the left. By defining F as the integrated function of f, which describes our system physics, we find that the final value of the IVP on $[t_i, t_{i+1}]$ depending on t_i as well as the control and state parametrization on that interval reads

$$F_i := F(t_{i+1}, \boldsymbol{x}(t), \boldsymbol{u}(t)) \ . \tag{8.57}$$

This yields for each point in time $(t_1, \cdots, t_{n_x - 1})$ the terminal value of the previous interval F_i as well as the initial value of the next interval s_i. To therefore ensure a smooth continuous trajectory over all intervals, we apply additional continuity constraints of kind

$$F_i - s_{i+1} = 0 \ \forall \ i \in [0, \cdots, n_x - 1] \tag{8.58}$$

that enforce continuity at interval transitions. An illustration of the result of this can be found in Fig. 8.4 on the right. Should the number of intervals be 1, this method is referred to as a *single shooting* approach, because only a single shot over one interval is taken to solve the time-dependent optimal control problem. Direct Multiple Shooting takes one shot per interval and ensures smoothness by employing eq. (8.58) as additional constraints to the optimization procedure.

Discretization and treatment of constraints

To apply the previously defined path constraint types to a Direct Multiple Shooting approach, we evaluate them on the respective grid points of the unified state and control grid $G_u \cup G_x$, assuming that they will hold sufficiently well within the intervals between the grid points.

Point constraints are easiest treated by choosing the underlying grids G_u and G_x such that the existence of an appropriate grid point is ensured for each time point t_k for which a point constraint is defined, i.e.

$$\{t_k\} \subset G_u \cap G_x \ . \tag{8.59}$$

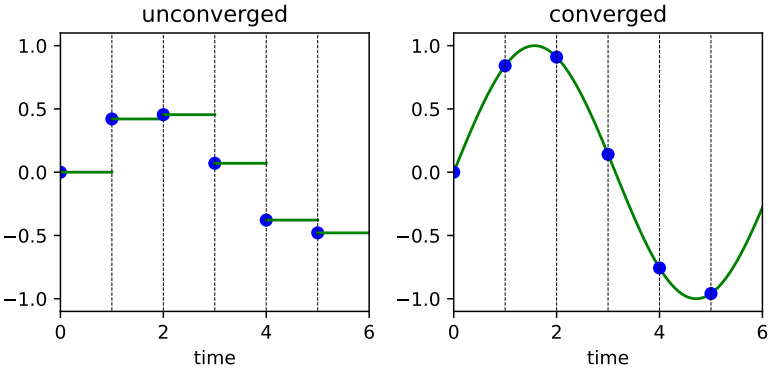

Figure 8.4.: State grid with computed state trajectories, without holding continuity conditions on the left and enforced continuity conditions on the right for six multiple shooting intervals. y-axis arbitraty for illustrative purposes only.

8.1.3. Discretized finite-dimensional constrained optimal control problem

Applying the aforementioned methodology, we can define the discretized, now finite-dimensional, constrained optimal control problem. We will apply an additional, but not necessary, constraint to simplify the problem: we will choose

$$G_u = G_x = G \ . \tag{8.60}$$

We can then define for our problem frame of work:

Definition 8. *We define G to be a shared grid for state and control variables. We furthermore define*

$$t \in G \tag{8.61}$$

and

$$\theta := (\boldsymbol{s}_0, \boldsymbol{c}_0, \cdots, \boldsymbol{s}_n, \boldsymbol{c}_n, \boldsymbol{p}) \ . \tag{8.62}$$

We can then define the discretized finite-dimensional constrained optimization problem as

$$\begin{aligned}
\min_{\theta} \quad & \Phi(\theta) && \text{(8.63a)}\\
\textit{subject to} \quad & \dot{\boldsymbol{x}} = f(t, \boldsymbol{x}(t), \boldsymbol{u}(t)) && \text{(8.63b)}\\
& \boldsymbol{g}(\theta) = 0 && \text{(8.63c)}\\
& \boldsymbol{h}(\theta) \geq 0 \ . && \text{(8.63d)}
\end{aligned}$$

8.1.4. Solving the discretized optimal control problem

A common way of solving problems of kind (8.63) is Sequential Quadratic Programming (SQP), which naturally emerges by the application of Newton's method to the Karush-Kuhn-Tucker (KKT) conditions that constitute the necessary conditions for optimality of nonlinear constrained optimization problems. We will now briefly review the foundational concepts of SQP, for an in-depth discussion of SQP we refer to [99].

Fundamentally, SQP iteratively solves nonlinear problems by generating an iterative sequence

$$\xi_{i+1} = \xi_k + \alpha\delta\xi \tag{8.64}$$

with $\delta\xi$ describing the direction of an iterative step and $\alpha \in \mathbb{R}^{>0}$ constituting the step length. Said steplength is obtained via line search, whereas the step direction is obtained by quadratically approximating (8.63), such that we obtain a new optimization problem that reads

$$\min_{\delta\xi_i} \quad \nabla\Phi(\xi_i)^T\delta\xi + \frac{1}{2}\delta\xi_i^T H_i \delta\xi_i \tag{8.65a}$$

$$\text{subject to} \quad \boldsymbol{g}(\xi_i) + \nabla\boldsymbol{g}(\xi_i)^T\delta\xi_i = 0 \tag{8.65b}$$

$$\boldsymbol{h}(\xi_i) + \nabla\boldsymbol{h}(\xi_i)^T\delta\xi_i \geq 0 \tag{8.65c}$$

with H_i denoting the Hessian of the Lagrangian. For more efficient solving of the aforementioned problem, H_i can be approximated by means of the BFGS or Gauss-Newton method or alternative approximations, the latter of which is particularly suited to least-squares objective functions. For further enhancements, the Jacobians $\nabla\boldsymbol{g}(\xi_i)$ and $\nabla\boldsymbol{h}(\xi_i)$ can be condensed by exploiting their high level of sparsity and structure, condensing these problems to smaller problems that only exhibit the minimal amount of variables necessary and compactify the remaining structure within a smaller problem that is typically more efficient to solve.

8.2. Reconstruction as an optimal control problem

To apply this method to our underlying problem, we will first formulate the reconstruction problem that will serve as our objective by defining the squared residual between the reference measurement $\hat{s}$ and the model prediction s for every sensor, yielding an overall objective function of

$$\Phi = \int_0^T \sum_{i\in I} \left(s_i - \hat{s}_i\right)^2 + \gamma\ddot{q}(t)^2 \mathrm{dt} \tag{8.66}$$

that should be minimal for an optimal reconstruction of the reference measurement for a given set of sensors $\{s_i\}_{i\in I}$, with γ being the strength of the regularization factor

$$\int_0^T \gamma\ddot{q}(t)^2 \mathrm{dt}\ , \tag{8.67}$$

which is chosen such that in case of non-uniqueness of the global optimum that motion is preferred that execerts the least amount of force. In contrast to a frame-by-frame reconstruction, this objective function allows to have a higher residual in some parts of the motion if that yields significantly favorable conditions for the residual in other parts of the motion. While the absolute difference would allow equal trading of residuals over the course of the motion, this is no longer the case for higher order polynomials, as the penalty is squared, increasing the impact of higher residuals and relatively reducing the impact of lower residuals over the different nodes.

These sensor readings, point accelerations and point velocities, can be computed based on a given set of generalized coordinates q and its derivatives for a rigid body kinematic tree model M by computing the kinematic chain of the associated rigid body parts and the offset of the point from the part's origin. This allows the simulation of IMU recorded data for a known reference motion $\vec{q}$ as well as its derivatives $\dot{\vec{q}}$ and $\ddot{\vec{q}}$, allowing the extraction of both acceleration as well as angular velocity for any given point that we assume an IMU to be placed in. We will

henceforth define an IMU at position p in the kinematic tree as $s(p, q, \dot{q}, \ddot{q})$, which we will, for ease of notation, express as $s(p, q)$ in the following, to derive a more detailed objective function for a given set of fixed IMUs $\{s(p_i, q)\}_{i \in I}$ as well as a reference motion expressed in its model coordinates $\hat{q}(t)$ as

$$\Phi = \int_0^T \sum_{i \in I} \left(s(p_i, q(t)) - \hat{s}_i\right)^2 + \gamma \ddot{q}(t)^2 \mathrm{dt} \tag{8.68}$$

with IMUs s_i located at the exact positions $\{p_i\}_{i \in I}$ that are also occupied by the reference IMUs $\hat{s}_i$.

We identify that the residual of Φ is dependent on the set of time dependent generalized coordinates $q(t)$ of M, and remembering our shorthand notation, also dependent on its derivatives $\dot{q}(t)$ and $\ddot{q}(t)$. To find the optimal set of parameters, we will formulate an optimal control problem based on the aforementioned objective function Φ and we will identify the differential state vector for this optimal control problem to be

$$\boldsymbol{x} = \begin{pmatrix} q \\ \dot{q} \end{pmatrix} \tag{8.69}$$

and consequentially, we will define its derivative as

$$\dot{\boldsymbol{x}} = f = \frac{d}{dt} \begin{pmatrix} q \\ \dot{q} \end{pmatrix} = \begin{pmatrix} \dot{q} \\ \ddot{q} \end{pmatrix} \tag{8.70}$$

which we identify in terms of OCP notation with an equation of kind

$$f = \frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ u \end{pmatrix} \ , \tag{8.71}$$

allowing us to express the state $(x_1, x_2) \in \mathbb{R}^{2n}$ in full dependence on the controls $u \in \mathbb{R}^n$, halving the effective degrees of freedom via the formulation of the model derivative. We can now solve for a time-dependent approximation of our generalized coordinates in dependence of their second derivative by employing Direct Multiple Shooting as outlined in the previous section.

While the resulting trajectory would be possible with the model we have specified, there is a high likelihood that it would still be impossible to perform for a human without severe joint damage, as we have not yet employed any physiological limits to the motion of the joints, allowing arbitrary overstretching. To prevent this and to retrieve a physiologically feasible solution, we will subject our optimal control problem to additional constraints: Firstly, we will subject the Pelvis body by constraining each of its degrees of freedom to $[0, 0]$ via additional equality constraints, effectively fixing it in place and serving as an anchor point. Practically, this can be achieved by both rewriting the motion trajectory relative to the anchor point, allowing for easier handling for the conceptual proposal of the following optimization procedure. The exact physical joint limits for the model can be found in Tab. 8.2. Joints that were not physically limited either due to physiological or reasons or conceptual simplifications in the model that do not allow for an easy transfer of physiological limits to the model were generally constrained to $[-20, 20]$, which was verified to be more than sufficient for the used reference trajectory, while also ensuring that the optimization area remained somewhat constrained, accelerating convergence. If a base motion does not allow such a comfortable estimate, these degrees of freedom can also be left unconstrained, at the cost of potentially slower convergence.

These joint constraints take the form of additional inequality constraints that get added to the optimal control formulation above with the specified upper and lower limits $l_{j,u}$ and $l_{j,l}$,

base body name	rot. x	rot. y	rot. z
Pelvis	$[0, 0]$	$[0, 0]$	$[0, 0]$
L5_L3	$[-\frac{\pi}{4}, \frac{\pi}{4}]$	$[-\frac{\pi}{4}, \frac{\pi}{4}]$	$[-\frac{\pi}{4}, \frac{\pi}{4}]$
T12_T8_Shoulders		$[-\frac{\pi}{4}, \frac{\pi}{4}]$	$[-\frac{\pi}{4}, \frac{\pi}{4}]$
Head_Neck		$[-\pi, \pi]$	n.a.
RightUpperArm	$[-20, 20]$	$[-20, 20]$	$[-20, 20]$
RightForeArm		$[-\pi, 0]$	$[-\pi, \pi]$
RightHand	$[-\pi, \pi]$		
LeftUpperArm	$[-20, 20]$	$[-20, 20]$	$[-20, 20]$
LeftForeArm		$[-\pi, 0]$	$[-\pi, \pi]$
LeftHand	$[-\pi, \pi]$		

Table 8.2.: List of bodies and their rotational joint limits. Empty fields correspond to non-existent degrees of freedom. Translational degrees of freedom are only applicable to the Pelvis body, which is fixed via limiting its rotational degrees of freedom as $[0, 0], [0, 0], [0, 0]$.

extending it to

$$\min_{\theta} \quad \sum_{t=t_0}^{t_m} \sum_{i \in I} \left(\left(s(p_i, q(t)) - \hat{s}_i \right)^2 + \gamma \ddot{q}_i(t)^2 \right) \tag{8.72a}$$

$$\text{subject to} \quad q_j = 0 \ \forall \ j \in [1, \ldots, 6] \tag{8.72b}$$

$$l_{j,u} - q_j(t_i) \leq 0 \ \forall \ i \in [1, \ldots, n], j \in [7, \ldots, n_{dof}] \tag{8.72c}$$

$$q_j(t_i) - l_{j,l} \leq 0 \ \forall \ i \in [1, \ldots, n], j \in [7, \ldots, n_{dof}] \ . \tag{8.72d}$$

with $m = 65$ being the number of shooting nodes used for this particular motion. In subsequent notations of the objective function $\gamma = 10^{-4}$ is always assumed if not stated otherwise.

With this reconstruction procedure at our fingertips, we can now employ this to make assessments of of reconstruction quality. To do so, we first need a reference motion in form of a model coordinate trajectory, i.e. for the set of q. This motion can either be extracted directly from the recording hardware, such as Xsens, or it can be obtained by employing the aforementioned reconstruction procedure with an appropriately high sensor density to allow for an accurate reconstruction. Alternatively, it can also be derived from other types of sensors, such as marker-based motion capture systems, whose reconstruction problem can be formulated analogously to the IMU sensor reconstruction outlined above.

Based on this reference recording $\hat{q}$, we can then derive the resulting sensor readings for IMUs placed at positions p_i by numerically deriving $\hat{q}$, if no derivatives were supplied, to obtain $\dot{\hat{q}}$ and $\ddot{\hat{q}}$ and forward simulating the motion based on the associated rigid-body kinematic tree model these trajectories were computed for. By applying this to the aforementioned reconstruction problem, we now rewrite the objective function that depended on fixed sensors into

$$\Phi_2(q) = \int_0^T \sum_{i \in I} \left(s(p_i, q(t)) - s(p_i, \hat{q}(t)) \right)^2 \mathrm{dt} + \gamma \int_0^T \ddot{q}(t)^2 \mathrm{dt} \ . \tag{8.73}$$

By doing so, this new objective function will yield an estimate of accuracy for a proposed reconstruction q and a set of sensors located at positions $\{p_i\}$ given a known trajectory $\hat{q}$ without the necessity to remeasure this trajectory for every set of positions $\{p_i\}$.

9. The impact of sensors, sensor properties and locations

9.1. The impact of noise

While we can reconstruct motion based on ideal sensor readings, as our proposed reconstruction assessment technique that computes forward-simulated sensor readings at arbitrary positions for any reference motion does allow for noise-free, ideal sensors, this can very obviously not be achieved in reality, as real hardware always has manufacturing tolerances that will result in measurement errors of some form. These errors can arise in the form of random noise due to limited tolerances or thermal noise on components, or through other components in the chain. Generally, they can be divided into two classes of errors:

- statistical errors (σ), which distribute randomly around a given measurement point
- systematic errors (μ), which systematically offset the measurement into a specific direction

To investigate the impact of both these kinds of measurement errors, we will augment our model of IMU sensors by adding a noise-term to the ideal reconstructed sensor reading $s(p, q, \dot{q}, \ddot{q})$, yielding

$$\tilde{s}(p, q, \dot{q}, \ddot{q}) = s(p, q, \dot{q}, \ddot{q}) + \mathcal{G}(\mu, \sigma) \tag{9.1}$$

which models both the statistical as well as the systematic error in one term in form of a Gaussian distribution with both error magnitudes as standard deviation as well as expectation value. Leveraging this modified parametrization of IMUs, we can rewrite our objective function again to

$$\Phi_2(q) = \int_0^T \sum_{i \in I} \left(\tilde{s}(p_i, q(t)) - \tilde{s}(p_i, \hat{q}(t)\right)^2 \mathrm{dt} + \gamma \int_0^T \ddot{q}(t)^2 \mathrm{dt} \ . \tag{9.2}$$

to obtain an optimal control based reconstruction problem that yields both an optimal reconstruction given a particular noise pattern as well as the residual objective alongside with it. Doing this, we can now vary the parameters for μ and σ and repeat the reconstruction procedure for these varying parameters to assess their detriment to the reconstruction. As this research is a continuation of the work in the Mitassist project, we will fundamentally assume that the noise properties of individual sensors can be assessed beforehand, i.e. we can measure μ and σ for the individual involved sensors prior to operation. We can now attempt a reconstruction procedure with varying levels of noise in four components:

- $\mu_a \in \Lambda_{\mu,a} \in \mathbb{R}^3$ quantifying the systematic offset in linear acceleration
- $\mu_\omega \in \Lambda_{\mu,\omega} \in \mathbb{R}^3$ quantifying the systematic offset in angular velocity
- $\sigma_a \in \Lambda_{\sigma,a} \in \mathbb{R}^3$ quantifying the random error on each measurement of linear acceleration
- $\sigma_\omega \in \Lambda_{\sigma,\omega} \in \mathbb{R}^3$ quantifying the random error on each measurement of angular velocity

We will define a noise configuration

$$\eta \in \Lambda_{\mu,a} \times \Lambda_{\mu,\omega} \times \Lambda_{\sigma,a} \times \Lambda_{\sigma,\omega} \tag{9.3}$$

with

$$\Lambda_{\mu,\nu} = \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix}, \mu_x, \mu_y, \mu_z \in \left[0, \frac{3}{8}, \frac{3}{4}, \frac{9}{8}, \frac{3}{2}\right] \ \forall\ \nu \in (a, \omega) \text{ s.t. } \mu_x = \mu_y = \mu_z \tag{9.4}$$

and

$$\Lambda_{\sigma,\nu} = \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix}, \sigma_x, \sigma_y, \sigma_z \in \left[0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right] \ \forall\ \nu \in (a, \omega) \text{ s.t. } \sigma_x = \sigma_y = \sigma_z \tag{9.5}$$

that defines the Gaussian distribution in eq. (9.1) that the noise for each measurement data point is drawn from, with

$$\mu = (\mu_a, \mu_\omega)^T \tag{9.6}$$

and

$$\sigma = (\sigma_a, \sigma_\omega)^T \ . \tag{9.7}$$

We will first discuss the impact of $\mu = (\mu_a, \mu_\omega)$ on the reconstruction, while keeping $\sigma_a = \sigma_\omega = 0$. Doing so will yield, for the parameter sweep specified in (9.4), 25 different reconstruction results. We find that these results are consistently reproducing the same reconstruction accuracy, no matter the choice of μ_a and μ_ω, the variance between these 25 different reconstructions is 5 magnitudes lower than the average between them. This is both intuitively expected, but also another strong quantitative evidence that with proper calibration of the measurement apparatus, systematic errors can be systematically driven into irrellevance, not posing a detriment to the reconstruction accuracy. It is, however, to be noted, that the necessary and potentially sensor-specific calibration procedures need to be performed before dispatching the system into use.

The more interesting question is posed by the impact of the statistical error, or sensor noise. Due to its statistical nature, no calibration procedure will allow to eliminate this error type, hence its impact onto the reconstruction needs to be assessed. We employ an equivalent approach to the assessment of the systematic error, sweeping the parameter space for σ_a and σ_ω according to (9.5) while keeping $\mu_a = \mu_\omega = 0$ for separation of effects (even if we have shown the negligible impact for μ). The results of this investigation are more impactful, they are depicted in Fig. 9.1, which depicts the average reconstruction error

$$\epsilon = \frac{1}{n_{dof} n_{shoot}} \sum_{i=1}^{n_{shoot}} \sum_{j=1}^{n_{dof}} \sqrt{(q_{i,j} - \hat{q}_{i,j})^2} \tag{9.8}$$

per shooting node and degree of freedom, normalized to a noise-free reconstruction error. Uncolored space depicts failed reconstructions, which clearly indicates that the chosen parameter sweep is challenging the capabilities of the reconstruction, indicating its behavior also for the realm of very strong noise levels. The first conclusion is, very obviously, that not every level of noise is viable for a successful reconstruction. This comes at no surprise as when the noise exceeds the signal, a reconstruction of the signal itself becomes increasingly difficult to the point of practical impossibility. However, for small enough noise levels, for this particular application $\sigma_a, \sigma_\omega \leq \frac{3}{4}$, we find a very stable area where all reconstruction attempts were successful. This area reaches up to about 150-fold increase in deviation from the reference trajectory compared to the noise-free reconstruction at $\sigma_i = 0$, up to a maximum increase to slightly over 400-fold of deviation compared to a noise-free reconstruction. At this increase, we find an average deviation of $0.12 rad$ or 7° per degree of freedom per shooting node to the reference trajectory set of joint

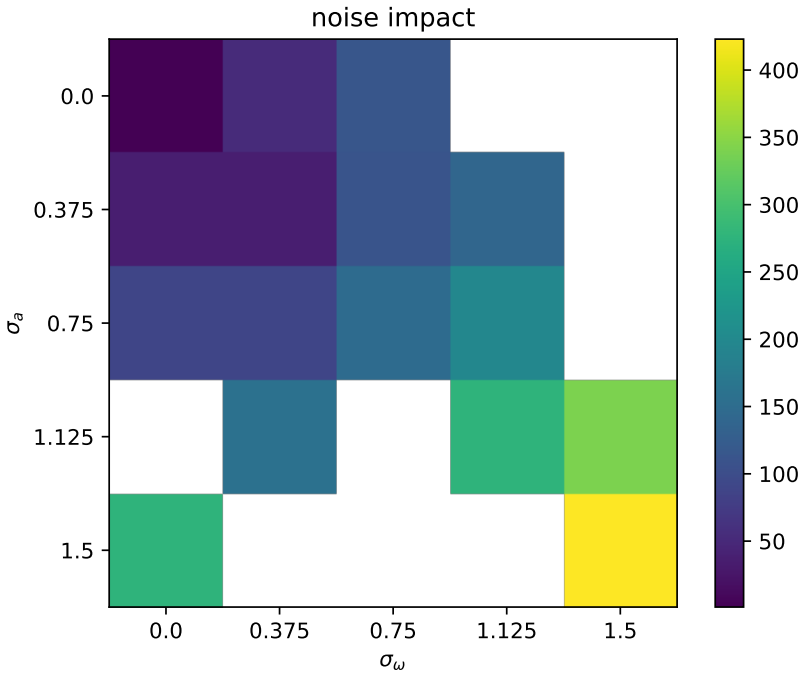

Figure 9.1.: Impact on reconstruction quality depending on noise level. White spots are reconstructions that did not succeed within 5000 SQP iterations, the colorbar represents the factor of increase in reconstruction error compared to the noise-free reconstruction.

angles, however, given the unstable environment this result is embedded into, this can be considered an already less indicative result. The consistent area that is enclosed with $\sigma_a = \sigma_\omega = 0.75$, in comparison to that, yields an average deviation of only $0.045rad$ or 2.5° deviation from the reference joint angles. To further investigate the behavior of the reconstruction degeneration with increasing levels of noise, we will take a look at the increase of reconstruction error on the diagonal of Fig. 9.1, which is separated out in Fig. 9.2. This yields an interesting insight: for the more stable area up to a noise level of $\frac{3}{4}$ for both components, we see an accelerating degradation in reconstruction quality, whereas the area above that exhibits a more linear behavior. Closer investigation has shown that, in contrast to the noise impact analysis in part 1 of this thesis neither an exponential function nor a low-order polynomial is an appropriate model to reproduce this behavior. Hence, the noise-acceleration curve is left as an interesting further direction to see if the noise-behavior can be modeled and appropriately predicted. Instead of investigating this further, we will conclude the demonstration of the capability of assessment of the impact of noise that this method allows, we will now continue with focusing on assessing the impact and relevance of individual sensors onto our reconstruction problem.

9.2. Assessing sensor relevance and location

After discussing the impact of sensor-specific properties, we will now continue to investigate the impact of sensor *configurations*, that is a set of multiple sensors with fixed or non-fixed locations and their impact on the configuration in three forms: We will discuss the relevance of individual

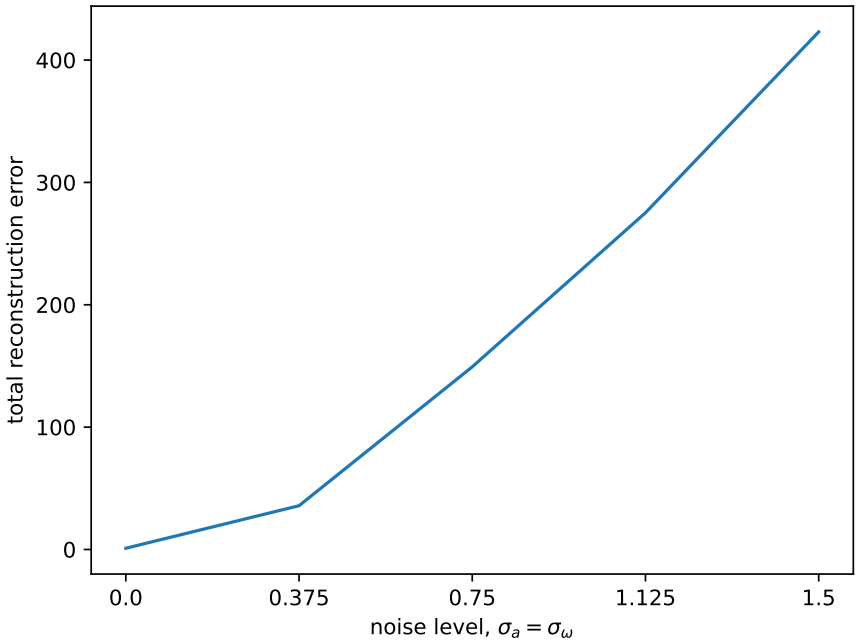

Figure 9.2.: Impact on reconstruction quality depending on noise level, relative to noise-free reconstruction.

sensors for both the overall reconstruction error as well as their impact on individual degrees of freedom. Further, we will briefly touch on the question of reconstruction accuracy when a reference configuration is reduced by multiple sensors at the same time. Eventually we will propose an approach to assess and optimize specific locations for multiple individual sensors.

9.2.1. Investigating the relevance of individual sensors for reconstruction

After showing how our proposed method is capable of assessing the impact of noise on the reconstruction procedure, we will now shift focus to assessing the relevance of individual sensors to the reconstruction problem to identify sensors whose absence does have a negligible impact onto the reconstruction result and which sensors exhibit a high relevance for said result. To do so we will again use the objective function (9.9) in its regularized form as described. The base IMU set that we are applying is placing one IMU on every single segment, except for the upper torso segment that represents the shoulder area as well as the upper part of the torso. Due to our sensor locations being derived from the Xsens [86] sensor layout, this segment is equipped with three IMUs, one on each shoulder and one on the sternum. They are named *T12_T8_Shoulders_A*, *T12_T8_Shoulders_B* and *T12_T8_Shoulders_C* respectively in the presented figures. To assess the relevance of these IMUs we will now modify the set of sensors $\{p_i\}_{i \in I}$ to obtain twelve different sets of sensors $\{p_i, j\}$ with each set missing the jth sensor. With these reduced sensor sets, we can then construct twelve reconstruction problems, each with

the objective

$$\Phi_{2,j}(q) = \int_0^T \sum_{i \in I} \left(s(p_{i,j}, q(t)) - s(p_{i,j}, \hat{q}(t))\right)^2 \mathrm{dt} + \gamma \int_0^T \ddot{q}(t)^2 \mathrm{dt} \tag{9.9}$$

which we set as an objective function for a reconstructive optimal control problem. We then assess the quality of the reconstruction, both in terms of absolute divergence from the reference trajectory as well as the individual contributions of the different degrees of freedom to this deviation, the result of which is depicted in Fig. 9.3. Multiple findings are immediately striking, some of which are more intuitive than others: Firstly, we find that both hands as well as the head exhibit substantial deviations from the reference trajectory, which is unsurprising due to all three segments being end-effectors, which means they are leaves of the kinematic tree. Consequentially, if they are not constrained by an IMU, their degrees of freedom can take arbitrary values without impacting the rest of the kinematic tree, and, with an accordingly reduced sensor set, without impacting the objective function used for reconstruction, which is defined solely based on the IMU measurement data, with the exception of the regularization term, which is particularly relevant in these cases for guaranteeing a unique solution, otherwise every value for the wrist joints and neck joints would yield the same value for the objective function, removing any tangible gradient and producing an optimization plateau. This also explains why the deviations in these degrees of freedom aren't higher, because the regularization term happens to be a somewhat good approximation to the actual motion of the wrist and neck joints for our particular exemplary motion as there was not much motion in either of the three, the dominant motion was in the arms. Hence, even larger deviations in the wrist and neck joints would be as entirely unsurprising as physically unindicative for different underlying motions.

Secondly, some IMUs have a negligible impact on the reconstruction quality. This is unsurprising for the IMUs located on the *T12_T8_Shoulders* segment, as their base segment is equipped with three of them. Dropping one, particularly without substantial amounts of noise, plausibly does not cause any substantial regression with respect to the reconstruction accuracy. A similar argument can be made for the Pelvis segment, which constitutes the root of the kinematic tree and is hence, as previously described, manually fixed in position to anchor the kinematic tree model for the development of this method. Hence, its potential to deviate from the reference trajectory is purposefully slimmed down. More interesting in this endeavour are the forearm segments, both left and right, because they are neither hand-constrained nor do they have additional IMUs that would address the missing IMU on the segment. Yet, we find negligible deviations from the reference trajectory also for these segments, which is a strong indicator that they are sufficiently constrained in their position and orientation by the neighbouring segments, namely the attached hand and upper arm, to which they are connected by three degrees of freedom for the joints, two to the upper arm and another in the wrist joint. We see that this constraints the forearm enough so that its trajectory is fully determined by the motion of the hand and upper arm. This is in stark contrast to leaving out the IMUs on the upper arms, which will result in substantial higher reconstruction errors. To understand how this comes to place we must first realize that in contrast to the end effectors, both the forearm and upper arm segments are embedded into the kinematic tree. So theoretically, they are defined by their surrounding segments. While that works well for the forearms, which are constraint by three degrees of freedom, it does less so for the upper arm segments that are constrained by five degrees of freedom, two in the elbow and three in the shoulder. One explanation for this is the alignment of two equally oriented degrees of freedom in the two joints. This scenario can be observed for the L5_L3 segment, which constitutes the center segment of the torso and is connected to both the pelvis as well as the T12_T8_Shoulders segment via a rotational joint in Z-axis. Hence, we can observe that when omitting the IMU on L5_L3, we find an increase in deviation from the reference

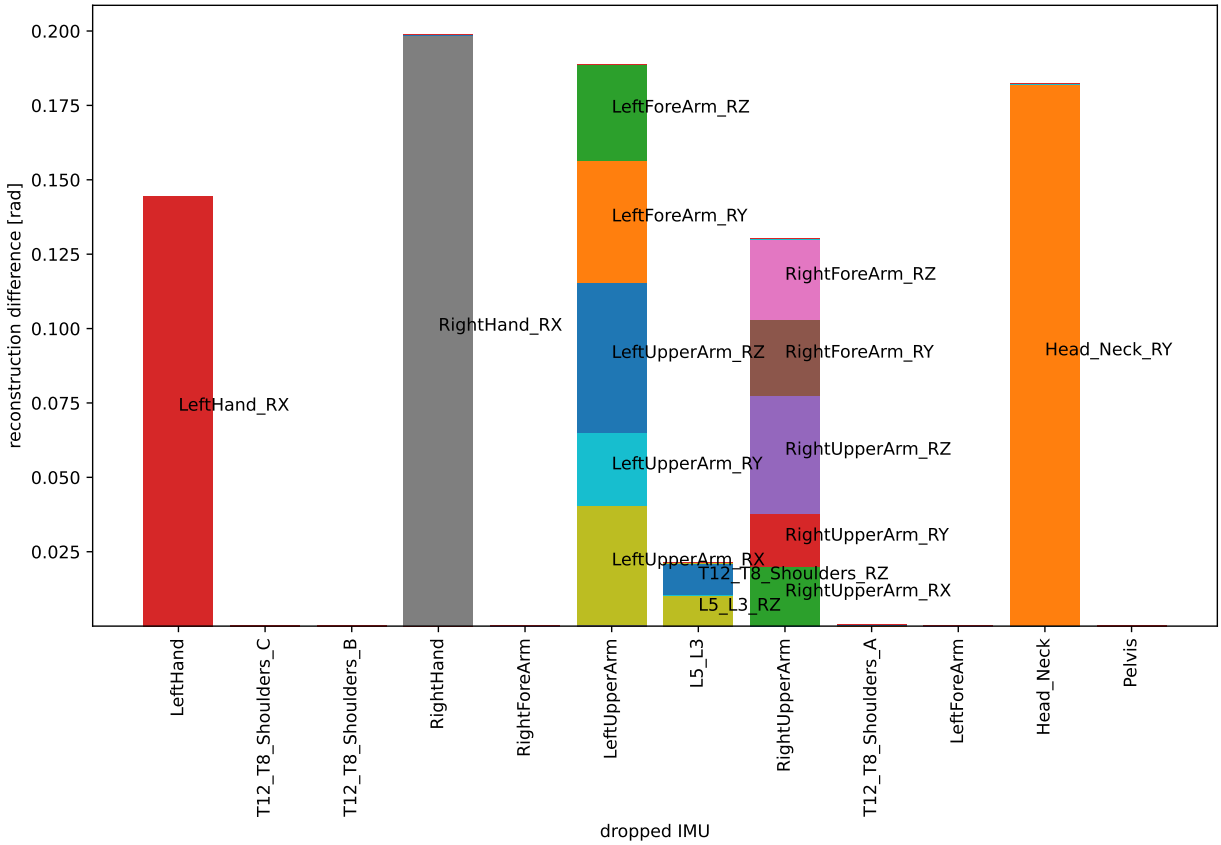

Figure 9.3.: Impact on reconstruction quality depending on noise level. On the x-axis we find the dropped IMU and the y-axis depicts the absolute differences from the reference trajectory per degree of freedom, averaged per shooting node (compare eq. (9.8)).

trajectory that equally distributes over the two aforementioned rotational degrees of freedom. This observation, however, does not translate to the forearm segments. We do find errors in the rotational Z component of both adjacent joints, which hints at the effect visible for L5_L3 at play, but we, firstly, see that the deviation is not equal in both of these degrees of freedom, and, secondly, we find the other degrees of freedom also deviating from the reference, an effect that was noticeably absent in L5_L3. However, we can identify that in the forearm we have in fact two kinds of redundant joints: We do have rotational degrees in the Z component, but also in the Y component. This leads to the emergence of a redundancy plane in the Z-Y-component that allows comparatively free motion of the upper arm segment, which explains why all these degrees of freedom are engaged in deviation from the reference trajectory. However, this will inevitably change the motion and consequentially change the objective function terms related to the IMUs on forearms and hands, which are corrected by engaging the free degree of freedom left in the X component of the shoulder. Both the L5_L3 and particularly the upper arm segments are strong indicators that segments that have degrees of freedom in the same component in multiple connections in the kinematic tree need to be tightly controlled and should therefore have a high priority when assigning sensors. This also yields an explanation why this effect is

entirely absent for the forearm: The three degrees of freedom affecting the forearm are spread over all three different axes, X in the wrist and Y and Z in the elbow. As a consequence, we do not have an axis redundancy that would allow for free rotation as the surrounding segments can independently constrain all three degrees of freedom related to the forearm. Hence, we can already conclude that IMUs are most effective in reducing deviations from the true motion when predominantly used to break redundancies in degrees of freedom within a model.

We can repeat this procedure and drop two IMUs at the same time and then perform a reconstruction, the overall residual of which can be seen in Fig. 9.4. While the overall additional insight is limited, it reinforces some of the findings of the previous analysis, in particular the high deviations from the reference in reconstructions that have an increased amount or length of end effectors, such as when hands and the corresponding forearm or multiple individual end effectors are not tracked together. Particularly the combination of missing hand and corresponding forearm cause the highest deviations in reconstruction that we see, even with tracking of the model joint angles only. Given that this effectively yields free, unconstrained lower arm movement, tracking the joint angles will even underestimate the absolute positional deviation of the motion if the effective untracked branch of the kinematic tree contains an end effector. Furthermore, we do see a general trend that one missing sensor yields a more accurate reconstruction than two missing sensors, which is intuitively plausible. We, however, also observe a few cases where the single missing sensor does not yield the best reconstructive result, such as with only the right hand missing yields a slightly worse trajectory accuracy on the joint angles as when T12_T8_Shoulders_B is missing as well, even though the difference is marginal. However, it indicates that for an application like this, the number of sensors must not be chosen too small to avoid the risk of overfitting. This problem can be reduced particularly by constraining the problem via appropriate models, but should be kept in mind for future research. This is typically observed for additional sensors missing that are very strongly constrained themselves, for example by having additional sensors on the same segment as it is the case for the T12_T8_Shoulders segment that contains three sensors in the full configuration of our base set. Another observation that is clearly reinforced by this analysis is the sensitivity of the overall trajectory deviation from the reference to the kind of objective function. A reconstructive objective based on IMUs, or potentially other kinds of sensors for that matter that constitute an inverse problem, can depending on parametrization find substantially different values for the same objective as we find in Fig. 9.4, indicating that depending on the chosen sensor configuration up to two magnitudes in overall residual remain, which again highlights the importance of choosing a good sensor layout and the usefulness of our proposed method to assess such layouts beforehand to ensure maximum information acquisition and retention.

9.2.2. Increased dropout and the impact of different sensors on reconstruction

To gain further insight on the criticality of different sensors, we will increase the strain on the reconstruction problem by removing more IMUs and repeat the procedure. To do so we will construct new sensor sets $\{p_j\}$ as subsets of the full set P so that the sensor number is reduced by four, i.e.

$$\{p_i\}_{i\in I} \subset P \text{ s.t. } \#I = \#P - 4 \tag{9.10}$$

which we will use to perform a reconstruction as outlined in section 8.2. By taking all possible sensor subsets that are of cardinality 8, i.e. fulfill the constraint $\#I = \#P - 4$ of four IMUs dropped, we obtain a distribution of 495 reconstructions, with each sensor being in the set of missing IMUs for 165 of them, which constitute a distribution of reconstructions for every sensor where it is missing that we can assess. This approach also means that there is sample

base body name	average	median	standard deviation
Head_Neck	0.47	0.37	0.51
LeftHand	1.19	0.45	3.39
LeftForeArm	0.94	0.31	2.87
LeftUpperArm	0.93	0.31	2.87
RightHand	1.02	0.46	2.97
RightForeArm	0.82	0.35	2.36
RightUpperArm	0.86	0.36	2.61
T12_T8_Shoulders_A	0.37	0.24	0.94
T12_T8_Shoulders_B	0.34	0.25	0.74
T12_T8_Shoulders_C	0.31	0.24	0.52
L5_L3	0.4	0.27	0.95
Pelvis	0.38	0.24	1.08

Table 9.1.: Average and median deviation from distribution of reference joint angles per degree of freedom and shooting node after dropping four IMUs, for each IMU being in the dropped set, as well as their standard deviation.

overlap between the distributions, which should be kept in mind to treat these results with appropriate care when deriving insight from them. To do this, we compute both average and the median as well as the standard deviation of the resulting average reconstruction residuals per degree of freedom and shooting node, which are listed in Tab. 9.1 and allow a number of conclusions: Firstly, we find that different end effectors are of different relevance, namely a missing sensor on the Head_Neck has a substantially smaller impact on reconstructions than the hands, which does make perfect sense when we consider their location in the kinematic tree. Whereas the Head_Neck segment is connected to T12_T8_Shoulders, which is fixed in place by three sensors, the hands are connected to the forearm segment, which in turn is connected to the upper arm segment, all three of which are held in place by one sensor only. Also, we know from earlier investigations that particularly the upper arm is a segment with a high amount of connected degrees of freedom, including redundant ones, and consequentially has a high potential for reconstruction errors. This substantially increases the impact of the hand endeffectors being unconstrained by a sensor, especially when the other missing sensors affect the other two arm segments. We also find a substantial difference between average and median for the left as well as the right hand, strongly indicating that there are a number of high deviation results that constitute a bad reconstruction, an effect that is substantially smaller for the Head_Neck segment. This is reaffirmed by similar effects that can be observed for the forearm and upper arm segments, which exhibit this effect for very similar reasons, whereas the core body of the model consisting of the T12_T8_Shoulders, L5_L3 and Pelvis segments are substantially stronger constrained, the first of them in particular due to the redundant sensor arrangement, yielding the smallest average deviations from the reference joint angles and are among the segments with the smallest spread, only undercut by the Head_Neck segment which happens to exhibit a motion that is well captured by our regularization term. This again reinforces our conclusions from the last section that both the number of associated degrees of freedom as well as the distance from the root of the kinematic tree for this model or the anchor point for an arbitrary model are good indicators for closer constrainment by sensor systems to achieve a better capture of the desired objective. From all 495 simulations done with the reduced sensor sets, 8 did not converge to a solution within 5000 SQP steps. All non-converged results had an IMU missing on at least one upper arm segment, either left or right, in 7 of them the missing IMU on the upper arm segment was combined with a missing IMU on either the associated forearm segment or the hand segment

of the same arm, again indicating the cruciality of constraining high-degree-of-freedom segments, particularly those that have redundant joint axes in said degrees of freedom.

9.2.3. Towards segment-specific locations of sensors

The immediate next step after evaluating the relevance of the sensors itself is assessing the impact of their exact location and find the optimal location for a specific objective. For demonstrational purposes we will stick with the aforementioned motion to constitute our objective, on which we will now derive a location optimization routine. This routine will be conducted in two steps: Firstly, we need to setup a surface parametrization to reduce the effective number of direct free variables of the problem by $\frac{1}{3}$. This has already outlined in section 8.0.2. Hence, we can directly continue with the second part, formulating this problem as a two-level optimization problem that allows both the assessment and, as a direct consequence, the selection of an optimized layout of sensors for the specific problem at hand. While less relevant for the previous section as the positions of the sensors were static and hence describing the location via two vs three parameters does not make any relevant difference apart from bookkeeping, the optimization problem will directly benefit not only from the reduced set of free parameters, but also from the incorporation of location constraints right into the problem description, which is what allows for this dimensionality reduction.

Consequentially, to formulate the location optimization problem, we will first define a fixed set of sensors

$$p = \{p_i\}_{i \in I} \tag{9.11}$$

that contains a number $\#p$ of sensors at locations p_i. We want to find these p_i such that for the sensor set p the reconstruction is optimal. To assess this, we will again assume the existence of a reference trajectory that can exemplarily be obtained by a reconstruction problem as described in section 8.2 with a high-cardinality sensor set and then serve as the reference we can compute the deviation from against, as performed already in the last section for a static layout. We recall the optimization objective (8.73) from the previous section, which we now reformulate from a predefined set p of sensors to depending on the set of sensors used for reconstruction, i.e.

$$\Phi_3(q,p) = \int_0^T \sum_{i \in I} \left(s(p_i, q(t)) - s(p_i, \hat{q}(t)\right)^2 \mathrm{dt} + \gamma \int_0^T \ddot{q}(t)^2 \mathrm{dt} \ . \tag{9.12}$$

Doing so allows to formulate a two-level optimization problem that reads

$$\min_p \quad \Phi_3(q,p) \tag{9.13}$$

$$\text{s. t.} \quad \min_q \quad \Phi_3(q,p) \ . \tag{9.14}$$

These two optimization problems are of different categories. The lower one is the time-dependent reconstruction problem, which can be formulated as an optimal control problem and will be solved as it was done in the previous sections. The upper level problem is a discrete, not time dependent optimization problem, which however depends on the result of the fairly costly-to-evaluate optimal control problem. As a consequence of that, the generation of a generalized derivative $\frac{\partial \Phi_3}{\partial p}$, even if done numerically, is undesirably costly. For that reason, we the upper level problem could be solved with an algorithm like BOBYQA [76], which is a derivative-free iterative optimization procedure and is hence well suited for the solution of this kind of problem. From there, several different configurations can be explored:

We can perform a stability analysis by preinitializing the sensor locations with a specific configuration p_{start} we want to test. This can, for example, be the configuration proposed by Xsens. If this configuration is optimal or close to optimal, the resulting positions p_{opt} should be very close to the initial localizatons p_{start}. This allows to uncover individual non-optimal sensors or a general non-optimal layout.

We can also assess a layout's stability with regards to different noise profiles. We have discussed the sensitivity of the reconstruction problem to noise for a static configuration already in section 9.1. Under this two-level optimization problem, we can investigate not only the stability, but also the optimality of such a layout given different levels of noise.

We further can investigate an optimal resulting configuration from an initialization $p_0 = \{0, \cdots, 0\}$ and see if a potentially stable configuration would also be the result of a zero-initialized optimization. This could yield particular insight into the question of local optima and will be particularly interesting for the question of global optimization given a somewhat challenging error surface.

It could also be indicative to how the initialization should be chosen when the sensor layout optimality is assessed with less sensors. Two scenarios come into mind: either, the provided layout is stable under sensor dropout or the majority of sensors will shift location to compensate for the information loss caused by the removed sensor. Particularly the former result would paint a particularly positive picture for the initial layout with regards to resistance to technical defects, which preferably never happen, but can never be ruled out entirely. This can of course be recursively performed by removing more sensors from a configuration until major shifts in the layout are observed, indicating the lower limit the provided layout is initialized with.

Unfortunately, due to time constraints this part could not be completed in time for this thesis, but the investigation will be continued.

9.3. Summary

In this final part of this thesis we have proposed a method that allows the determination of an optimal sensor layout for a specific application. We have outlined an exemplary model problem in form of a motion reconstruction based on the framework of optimal control, which, while used for a kinematic reconstruction in our example, can easily be extended to more complex rigid-body models. As these are a popular approach for both biomechanical and robotic applications, including additional physical properties, which opens the door to a wide variety of application options. We have shown that our framework allows not only for the assessment of hardware parameters such as measurement noise and its impact onto the reconstruction, allowing for a pre-selection of components without the necessity of building physical prototypes and exploring their practicability in the field, but also allows for estimating the relevance of different sensors in a predefined setup. We have shown that this method allows both the assessment of relevance with respect to the overall reconstruction quality as well as the impact on individual degrees of freedom of the underlying model, allowing for an informed decision on the corresponding sensor's importance considering whether these degrees of freedom need to be captured accurately or if they happen to be less relevant. We have further derived an exemplary surface parametrization of such a model that uses model geometry to reduce the dimensionality of the overall search space and at the same time allows for a free-floating optimization of locations. We have established a two-level problem to implement such a free-floating optimization based on the aforementioned reconstruction problem. Given the discontinuities of the surfaces of the model, this needs to be carefully investigated to avoid the creation of singularities. This was not not possible within the time limits of this thesis project, but the goal appears to be clearly achievable by appropriate parametrization, allowing for a fully free relocation of sensors on a model surface. This would

then finally yield the optimal sensor layout for an arbitrary problem that allows to build cost-effective and operationally simple sensor systems for a vast number of different applications and advancing the technological development of technologies and research based on it.

This would likely be beneficial for the optimization routine with regards to performance, but also has certain drawbacks, the most prominent one is the higher level of integration between the two levels, which could be detrimental to pluggability of the lower level.

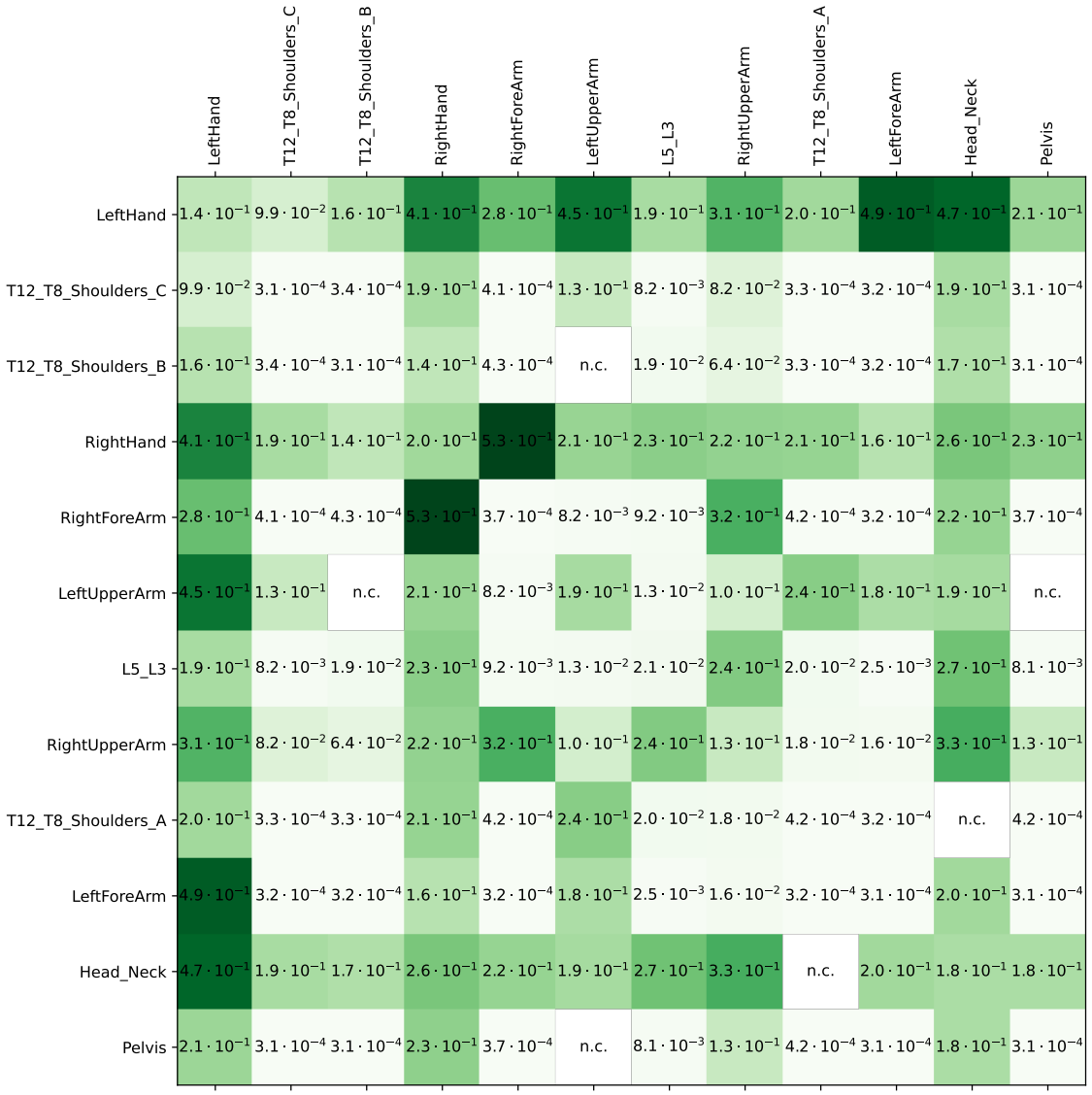

Figure 9.4.: Deviation in Radiant (see eq. (9.8)) from reference trajectory, averaged over all effective degrees of freedom and all shooting nodes, after reconstruction depending on on two IMUs (or one IMU on the diagonal for reference) removed from the overall IMU set per shooting node and degree of freedom. “n.c.” refers to no convergence within 5000 SQP steps.

10. Conclusion and outlook

In this thesis, we have explored the question of motion classification in various different ways: We have started with a classification of *direct motion* in the form of hand gestures based on their muscle signatures. We have then continued to use motion as a *predictor* for aversive cognitive states in the context of major depressive disorder. Finally, we have investigated the question of assessing the *quality* of a sensor configuration for a specific purpose and ways to *improve and optimize* such a configuration.

10.1. Key findings of this work

In the **first part**, we have investigated the direct detection of motion by means of muscle signatures, and have investigated how variations in the time axis of execution affect the overall classification. We could show why this classification approach is successful and could define criteria that make it successful. We further showed that the sample size required for a well-performing classification procedure is comparatively low. In addition to that, improvements to the computational complexity of our method by means of data condensation still yields similar classification performance as the full dataset at significantly lower computational cost.

We have then proceeded in the **second part** to identifying averse cognitive states in the context of major depressive disorder. We have demonstrated, for the first time, that motion can serve as a predictor to differentiate between states of acute cognitive aversion and the lack thereof. We have further demonstrated that this is the case for patient-individual as well as a shared dataset. The individual dataset unsurprisingly exhibited higher accuracies, but the shared dataset still yielded noticeable improvements, indicating that there might be cross-patient structure in the data. We also demonstrated how different components of this method contribute to the resulting accuracies in terms of correctly identifying averse cognitive states or correctly rejecting non-averse cognitive states via investigating sensitivity, specificity and precision. However, the results need to be read with a bit of caution as it should be verified how reliable the participants' self assessments are, for example by comparing classification accuracies of samples with varying proximity to the survey point. If the accuracy of more distant samples turns out to be lower than the accuracy of samples closer to the time point of the survey it could be an indication that participant judgement is less reliable and a security margin should be applied to the data considered for evaluation. This could either be by only using the second half of the labeled data or recursively assessing the plausibility of the participant's self-assessment, if a good measure for such a recursive assessment can be derived.

Furthermore the approach we have proposed has a substantial number of degrees of freedom. While we have a substantial amount of acquired patient data for reference for the kind of study that was possible to perform, for this specific problem the amount of data remains limited. This increases uncertainty in the results, which contributes to a higher spread in resulting accuracies. This can either be solved by increasing the amount of data, or by reducing the number of degrees of freedom by selecting the most prominent configurations and eliminating degrees of freedom that we have already discussed not to be beneficial.

In the final and **third part**, we have proposed an approach that allows to assess the importance of specific IMU-sensors for the reconstruction of an upper-body motion and derived from there a procedure that will eventually allow the optimization of sensor layouts for a specific reconstruction problem. We have shown that specific segments are in higher need for tracking by sensors, particularly if they are relevant endeffectors or when they have coinciding degrees of freedom that need to be tracked to avoid rotation in the free axis. We have finally proposed a method that also allows the assessment and optimization of sensor locations, and have outlined several scenarios in which this method can be used, from stability analyses to the investigation of the impact of noise onto the suitability of the specific sensor layout in addition to the penalty to the reconstruction itself. Due to time constraints, this last part could not be demonstrated in practice yet and hence remains conceptual for the moment. The method is kept such that the underlying model can be exchanged depending on application requirement, allowing for great flexibility. One application that comes to mind is the assessment of relevance of sensors for the detection of aversive cognitive states in depressive disorder or other mental disorders, where such a method could provide insight into the relevance of different components onto the overall classification accuracy.

10.2. Future directions of research

While work conducted in this thesis spans from direct classification of signals in form of EMG signatures to a more indirect approach in the form of detecting aversive states based on specific motion characteristics, all of them offer several directions to continue research:

The question of EMG signatures for hand gestures was, up to now, primarily investigated in an in-depth analysis of two individual participants, and while the approach itself does not depend on the number of subjects, it would be highly interesting to see how consistent these results are. As we could show that this approach is noticeably more stable for strong signals, more participants would allow for a better assessment the exact impact of the strength of the general muscle signal response. Furthermore, an increased number of participants would allow revisiting the question of transferability which we, based on the data we had, concluded negatively. However, it cannot and should not be ruled out that a degree of transfer might be possible under certain criteria. These criteria should be investigated with a dataset containing more participants, with potentially less samples per participant, which also eases acquisition.

Another interesting aspect is a variation of the set of gestures considered, and potentially establishing a criterion that allows to automatically define a subset of gestures from a larger pool with maximum separability. This way, a user could provide a superset of gestures from which a reduced number is automatically chosen with maximum separability.

Also, testing the proposed procedure with less sensors could yield additional insight into the stability requirements. If the method would yield similar classification accuracies with less sensors, this would simplify manufacturing hardware for this application and would also reduce the computational burden on the hardware. Similar to this, it might be interesting to test different locations of sensors. The Myo-Armband, the acquisition hardware for this specific investigation, uses a set of sEMG-electrodes around the forearm. While this is most certainly not a bad location, it is also not muscle specific, which is both an advantage in terms of simple operation for the user, but a muscle-specific placement of such electrodes as many studies do could yield additional improvements, albeit at the expense of simplicity in operation. Connected to this is the question of wet versus dry electrodes. Again relevant for the simplicity of operation, wet electrodes are the de facto standard in motion capture applications because a highly conductive electrode will collect more charges than a dry electrode, particularly if the electrode's location

is shaved prior to electrode application. This would come as a substantial reduction in user-autonomous operability of such hardware, but would nevertheless yield interesting insight into the impact that dryness and potential hair between electrode and skin have in practice. Finally, an application of this approach to a continuous datastream has not been performed yet under a controlled environment and would be the direct next step. However, all necessary parts, in particular the rejection logic for gesture classification, have been laid out to implement this.

The classification of patient data with regards to the identification of aversive cognitive states in the context of major depressive disorder similarly yields several further directions to take: Firstly, while the number of patients included into this study are high based on the difficulty of the acquisition procedure and the limited resources present, it would allow for a stronger analysis if more patient datasets would be available that yield enough data for the individual-focused analysis, providing a stronger statistical base on which our results can be verified. Equally, this will strengthen the cross-patient results, as the statistical foundation is broader, addressing the highly individualistic nature of the disorder that we have observed. Secondly, we have optimized the configurations for maximum accuracy. In a practical setting, however, other objectives might be more desirable, such as a preference for high sensitivities in the classification configurations to ensure dysfunctional states are identified. At the same time, an identifier with high sensitivity is easy to make by simply identifying everything as relevant, hence such an objective would have to be balanced with either an accuracy- or a precision-term, creating a composite objective function that could increase flexibility and better serve specific treatment intentions at the discretion of the therapist. Thirdly, this approach, already yielding very good results for major depressive disorder, is not limited to this specific type of disorder. Broadening the scope to other disorders, such as obsessive compulsive disorder or anxiety disorder, will increase the impact on therapy assistance at almost no cost, as the method is deliberately constructed to be disorder-agnostic, possibly being useful even outside the field of psychotherapeutic assistance or even medicine in total. Fourthly, this investigation has focused in using motion as a predictor. The underlying hardware provides more types of data, such as heart rate variability, that could yield additional insight and even better identifiability of aversive states and should without a doubt be included in the future to further improve the already very good results we have presented.

The last part investigating the optimality of sensor locations, also provides a number of directions to proceed. First and foremost, the discussed oddities with the location-optimization routine need to be addressed and the root cause identified and corrected. Once that is done, segment-switches need to be implemented and a way needs to be found to deal with potentially arising discontinuities at segment junctions. Afterwards, this will allow to find optimal sensor layouts not only for the specific motion we used in this thesis, but to other motions as well, or entire motion suites that can be comprised of all motions relevant for a specific study and then finding an optimal layout for sensors to measure and reconstruct all motions in the set with minimal sensor usage and maximum reconstruction quality. This will be of particular interest when the number of sensors is reduced to see what is the minimal set of sensors where reconstruction is still viable.

This last question is also relevant for the second part of this thesis, as the number of sensors is relevant to both the amount of data obtained for analysis, but at the same time more sensors could mean more components to operate, which will reduce patient willingness to participate in an assistive based on such a device as well as increase errors in handling, as the idea revolves around patient-autonomous monitoring of cognitive state.

An approach such as our proposed sensor location optimization can be of use to assess the relevance of different sensors for such a problem, potentially indicating that the number of sensors can be reduced at a specific point or must not be reduced in another location, aiding in deciding

how to trade simplicity with effectiveness and to identify associated pareto optima.

Bibliography

[1] Flachgipflig.svg. https://de.wikipedia.org/wiki/Datei:Flachgipflig.svg, 2009.

[2] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[3] M. R. Ahsan, M. I. Ibrahimy, and O. O. Khalifa. Electromygraphy (EMG) signal based hand gesture recognition using artificial neural network (ANN). In *2011 4th International Conference on Mechatronics (ICOM)*, pages 1–6, May 2011.

[4] G. Batista, E. Keogh, O. Tataw, and V. deSouza. CID: an efficient complexity-invariant distance measure for time series. *Data Mining and Knowledge Discovery*, 28(3):634–669, 2014.

[5] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[6] Diehl M. Kirches C. Mombaur K. Sager S. Bock, H.G. Optimierung bei gewöhnlichen differentialgleichungen. University Lecture, 2014.

[7] HG Bock, MM Diehl, DB Leineweber, and JP Schlöder. A direct multiple shooting method for real-time optimization of nonlinear dae processes. In *Nonlinear model predictive control*, pages 245–267. Springer, 2000.

[8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[10] BÄK, KBV, AWMF. S3-Leitlinie/Nationale VersorgungsLeitlinie (NVL) Unipolare Depression, 2. Auflage. 2015.

[11] K. F. Chung and K. C. Tso. Relationship between insomnia and pain in major depressive disorder: A sleep diary and actigraphy study. *Sleep Med*, 11(8):752–758, Sep 2010.

[12] R. Churchill, V. Hunot, R. Corney, M. Knapp, H. McGuire, A. Tylee, and S. Wessely. A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression. *Health Technol Assess*, 5(35):1–173, 2001.

[13] J. D. Cook, M. L. Prairie, and D. T. Plante. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *J Affect Disord*, 217:299–305, 08 2017.

[14] Francesca Cormack, Maggie McCue, Nick Taptiklis, Caroline Skirrow, Emilie Glazer, Elli Panagopoulos, Tempest A van Schaik, Ben Fehnert, James King, Jennifer H Barnett, et al. Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: longitudinal observational study. *JMIR mental health*, 6(11):e12814, 2019.

[15] P. Cover, T.; Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 1967.

[16] Renate de Jong-Meyer, Martin Hautzinger, Gerhard A Rudolf, and Wolfgang Strauß. Die überprüfung der wirksamkeit einer kombination von antidepressiva-und verhaltenstherapie bei endogen depressiven patienten: Varianzanalytische ergebnisse zu den haupt-und nebenkriterien des th. *Zeitschrift für Klinische Psychologie*, 1996.

[17] R. J. DeRubeis, L. A. Gelfand, T. Z. Tang, and A. D. Simons. Medications versus cognitive behavior therapy for severely depressed outpatients: mega-analysis of four randomized comparisons. *Am J Psychiatry*, 156(7):1007–1013, Jul 1999.

[18] S. Dimidjian, S. D. Hollon, K. S. Dobson, K. B. Schmaling, R. J. Kohlenberg, M. E. Addis, R. Gallop, J. B. McGlinchey, D. K. Markley, J. K. Gollan, D. C. Atkins, D. L. Dunner, and N. S. Jacobson. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *J Consult Clin Psychol*, 74(4):658–670, Aug 2006.

[19] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, 1(2):1542–1552, August 2008.

[20] Vanderplas J Metzen J.H. Lemaitre G. Dubourg, V. Gaussian Processes regression: basic introductory example. `https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_noisy_targets.html`.

[21] Mark Ebden. Gaussian processes: A quick introduction. *arXiv preprint arXiv:1505.02965*, 2015.

[22] R. Featherstone. Plucker basis vectors. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 1892–1897, 2006.

[23] Martin L Felis. Rbdl: an efficient rigid-body dynamics library using recursive algorithms. *Autonomous Robots*, 41(2):495–511, 2017.

[24] Martin Leonhard Felis. *Modeling emotional aspects in human locomotion.* PhD thesis, 2015.

[25] Anthony V Fiacco and Jiming Liu. On the stability of general convex programs under slater's condition and primal solution boundedness. *Optimization*, 32(4):291–299, 1995.

[26] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In Paul Vitányi, editor, *Computational Learning Theory*, pages 23–37, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

[27] Wolfgang Gaebel, Sandra Kowitz, Jürgen Fritze, and Jürgen Zielasek. Use of Health Care Services by People With Mental Illness. *Dtsch Arztebl International*, 110(47):799–808, 2013.

[28] François Petitjean; Alain Ketterlin; Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44, 2011.

[29] V. Gloaguen, J. Cottraux, M. Cucherat, and I. M. Blackburn. A meta-analysis of the effects of cognitive therapy in depressed patients. *J Affect Disord*, 49(1):59–72, Apr 1998.

[30] Omer Gold and Micha Sharir. Dynamic time warping and geometric edit distance: Breaking the quadratic barrier. *ACM Transactions on Algorithms (TALG)*, 14(4):1–17, 2018.

[31] T. Górecki and M. Łuczak. Non-isometric transforms in time series classification using DTW. *Knowledge-Based Systems*, 61:98–108, 2014.

[32] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. Learning time-series shapelets. In *proceedings of 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[33] D. Guijo-Rubio, P. Gutiérrez, R. Tavenard, and A. Bagnall. A hybrid approach to time series classification with shapelets. In *proceedings of Intelligent Data Engineering and Automated Learning*, volume 11871 of *Lecture Notes in Computer Science*, pages 137–144. 2019.

[34] D. Guijo-Rubio, P. Gutiérrez, R. Tavenard, and A. Bagnall. A hybrid approach to time series classification with shapelets. In *proceedings of Intelligent Data Engineering and Automated Learning*, volume 11871 of *Lecture Notes in Computer Science*, pages 137–144. 2019.

[35] G. E. Hardy, J. Cahill, W. B. Stiles, C. Ispan, N. Macaskill, and M. Barkham. Sudden gains in cognitive therapy for depression: a replication and extension. *J Consult Clin Psychol*, 73(1):59–67, Feb 2005.

[36] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[37] E. C. Harris and B. Barraclough. Suicide as an outcome for mental disorders. A meta-analysis. *Br J Psychiatry*, 170:205–228, Mar 1997.

[38] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[39] Martin Hautzinger and Renate de Jong-Meyer. Zwei multizenter-studien zur wirksamkeit von verhaltenstherapie, pharmakotherapie und deren kombination bei depressiven patienten: Einführung, rahmenbedingungen und aufgabenstellungen. *Zeitschrift für klinische Psychologie*, 1996.

[40] Martin Hautzinger, Renate de Jong-Meyer, Renate Treiber, and Gerhard A Rudolf. Wirksamkeit kognitiver verhaltenstherapie, pharmakotherapie und deren kombination bei nicht-endogenen, unipolaren depressionen. *Zeitschrift für Klinische Psychologie*, 1996.

[41] James L Hedlund and BW Vieweg. The hamilton rating scale for depression: a comprehensive review. *Journal of Operational Psychiatry*, 10(2):149–165, 1979.

[42] Rodolfo Hermans. Negative and positive skew diagrams. https://commons.wikimedia.org/w/index.php?curid=4567445, 2008.

[43] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[44] S. D. Hollon, R. J. DeRubeis, M. D. Evans, M. J. Wiemer, M. J. Garvey, W. M. Grove, and V. B. Tuason. Cognitive therapy and pharmacotherapy for depression. Singly and in combination. *Arch Gen Psychiatry*, 49(10):774–781, Oct 1992.

[45] S. D. Hollon, R. F. Muñoz, D. H. Barlow, W. R. Beardslee, C. C. Bell, G. Bernal, G. N. Clarke, L. P. Franciosi, A. E. Kazdin, L. Kohn, M. M. Linehan, J. C. Markowitz, D. J. Miklowitz, J. B. Persons, G. Niederehe, and D. Sommers. Psychosocial intervention development for the prevention and treatment of depression: promoting innovation and increasing access. *Biol Psychiatry*, 52(6):610–630, Sep 2002.

[46] A. Hoogerhoud, A. W. Hazewinkel, R. H. Reijntjens, I. M. van Vliet, M. S. van Noorden, G. J. Lammers, J. G. van Dijk, and E. J. Giltay. Short-term effects of electroconvulsive therapy on subjective and actigraphy-assessed sleep parameters in severely depressed inpatients. *Depress Res Treat*, 2015:764649, 2015.

[47] Roger A Horn and Charles R Johnson. Norms for vectors and matrices. *Matrix analysis*, 1990.

[48] L. Hou, J. Kwok, and J. Zurada. Efficient learning of timeseries shapelets. In *proceedings of 30th AAAI Conference on Artificial Intelligence*, 2016.

[49] Gan Huang, Dingguo Zhang, Xidian Zheng, and Xiangyang Zhu. An emg-based handwriting recognition through dynamic time warping. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 4902–4905, 2010.

[50] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[51] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[52] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[53] William B Johnson. Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26:189–206, 1984.

[54] I. Karlsson, P. Papapetrou, and H. Boström. Generalized random shapelet forests. *Data Mining and Knowledge Discovery*, 30(5):1053–1085, 2016.

[55] M. Kim; J. Lee; K. Kim. Enhancement of sEMG-based gesture classification using mahanobis distance metric. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 1117–1122, June 2016.

[56] Stephan; André Elisabeth Kim, Jonghwa; Mastnik. [ACM Press the 13th international conference - Gran Canaria, Spain (2008.01.13-2008.01.16)] Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08 - EMG-based hand gesture recognition for realtime biosignal interfacing. 2008.

[57] Eduardo Laber and Lucas Murtinho. Minimization of gini impurity: Np-completeness and approximation algorithm via connections with the k-means problem. *Electronic Notes in Theoretical Computer Science*, 346:567–576, 2019. The proceedings of Lagos 2019, the tenth Latin and American Algorithms, Graphs and Optimization Symposium (LAGOS 2019).

[58] Stephen Lake. Ending sales of myo, preparing for the future.

[59] Inc. LeapMotion. Leapmotion. `https://www.leapmotion.com/technology/`.

[60] Seunggyu Lee, Hyewon Kim, Mi Jin Park, and Hong Jin Jeon. Current advances in wearable devices and their sensors in patients with depression. *Frontiers in Psychiatry*, 12:672347, 2021.

[61] Daniel B Leineweber and Interdisziplin ares Zentrum. *Analyse und Restrukturierung eines Verfahrens zur direkten L osung von Optimal-Steuerungsproblemen (The Theory of MUSCOD in a Nutshell)*. PhD thesis, Diploma thesis, University of Heidelberg, 1995.

[62] Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, 2006.

[63] B. Lucas, A. Shifaz, C. Pelletier, L. O'Neill, N. Zaidi, B. Goethals, F. Petitjean, and G. Webb. Proximity forest: an effective and scalable distance-based classifier for time series. *Data Mining and Knowledge Discovery*, 33(3):607–635, 2019.

[64] C. McCall and W. V. McCall. Comparison of actigraphy with polysomnography and sleep logs in depressed insomniacs. *J Sleep Res*, 21(1):122–127, Feb 2012.

[65] G. E. Murphy, A. D. Simons, R. D. Wetzel, and P. J. Lustman. Cognitive therapy and pharmacotherapy. Singly and together in the treatment of depression. *Arch Gen Psychiatry*, 41(1):33–41, Jan 1984.

[66] Meinard Müller. *Information Retrieval for Music and Motion // Dynamic Time Warping*, volume 10.1007/978-3-540-74048-3. 2007.

[67] North. Myo armband. `https://support.getmyo.com`.

[68] JT O'Brien, P Gallagher, D Stow, N Hammerla, T Ploetz, M Firbank, C Ladha, K Ladha, D Jackson, Roisin McNaney, et al. A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychological medicine*, 47(1):93–102, 2017.

[69] World Health Organization. Icd-10 : international statistical classification of diseases and related health problems : tenth revision, 2004.

[70] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

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

[72] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry*, 11:584711, 2020.

[73] Ignacio Peis, Javier-David López-Moríñigo, M Mercedes Pérez-Rodríguez, Maria-Luisa Barrigón, Marta Ruiz-Gómez, Antonio Artés-Rodríguez, and Enrique Baca-García. Actigraphic recording of motor activity in depressed inpatients: a novel computational approach to prediction of clinical course and hospital discharge. *Scientific reports*, 10(1):1–11, 2020.

[74] M.; Huang T.S. Pengyu Hong, ; Turk. [IEEE Comput. Soc Fourth International Conference on Automatic Face and Gesture Recognition - Grenoble, France (28-30 March 2000)] Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580) - Gesture modeling and recognition using finite state machines. 2000.

[75] S. Pitou, F. Wu, A. Shafti, B. Michael, R. Stopforth, and M. Howard. Embroidered electrodes for control of affordable myoelectric prostheses. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1812–1817, May 2018.

[76] Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26, 2009.

[77] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[78] J Ross Quinlan. Program for machine learning. *C4. 5*, 1993.

[79] Nadine Raoux, Odile Benoit, Nicolas Dantchev, Pierre Denise, Bernard Franc, Jean-François Alliale, and Daniel Widlöcher. Circadian pattern of motor activity in major depressed patients undergoing antidepressant therapy: relationship between actigraphic measures and clinical course. *Psychiatry research*, 52(1):85–98, 1994.

[80] Jingjing; Yuan Junsong; Zhang Zhengyou Ren, Zhou; Meng. [ACM Press the 19th ACM international conference - Scottsdale, Arizona, USA (2011.11.28-2011.12.01)] Proceedings of the 19th ACM international conference on Multimedia - MM '11 - Robust hand gesture recognition with kinect sensor. 2011.

[81] A. Sandmeir, D. Schoenherr, U. Altmann, C. Nikendei, H. Schauenburg, and U. Dinger. Depression Severity Is Related to Less Gross Body Movement: A Motion Energy Analysis. *Psychopathology*, 54(2):106–112, 2021.

[82] Anna Sandmeir, Désirée Schoenherr, Uwe Altmann, Christoph Nikendei, Henning Schauenburg, and Ulrike Dinger. Depression severity is related to less gross body movement: a motion energy analysis. *Psychopathology*, 54(2):106–112, 2021.

[83] Lawrence K Saul and Sam T Roweis. An introduction to locally linear embedding. *http://www.cs.toronto.edu/~roweis/lle/publications.html*, 2000.

[84] P. Schäfer. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.

[85] P. Schäfer. Scalable time series classification. 30(5):1273–1298, 2016.

[86] Martin Schepers, Matteo Giuberti, Giovanni Bellusci, et al. Xsens mvn: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8), 2018.

[87] K. Schnell and S. C. Herpertz. Psychotherapy in psychiatry: the current situation and future directions in Germany. *Eur Arch Psychiatry Clin Neurosci*, 261 Suppl 2:S129–134, Nov 2011.

[88] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[89] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[90] H Strotzka. Psychotherapie: Grundlagen, verfahren. *Indikationen. München*, 1975.

[91] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[92] M. E. Thase, J. B. Greenhouse, E. Frank, C. F. Reynolds, P. A. Pilkonis, K. Hurley, V. Grochocinski, and D. J. Kupfer. Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Arch Gen Psychiatry*, 54(11):1009–1015, Nov 1997.

[93] Paul E. Utgoff. Id5: An incremental id3. In *ML*, 1988.

[94] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[95] Walber. Precision and recall. `https://commons.wikimedia.org/wiki/File:Precisionrecall.svg`, 2014.

[96] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364, 2019.

[97] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[98] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[99] Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

[100] V.; Kongqiao Wang; Jihai Yang Xu Zhang; Xiang Chen; Yun Li; Lantz. A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors. *IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans*, 41, 2011.

[101] M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. Dau, D. Silva, A. Mueen, and E. Keogh. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*, 32(1):83–123, 2018.

[102] J. Zakaria, A. Mueen, E. Keogh, and N. Young. Accelerating the discovery of unsupervised-shapelets. *Data mining and knowledge discovery*, 30(1):243–281, 2016.

[103] Harry Zhang. The optimality of naive bayes. *Aa*, 1(2):3, 2004.

[104] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19:4–12, April 2012.

[105] Danny Zhu. myo-raw data acquisition library. `https://github.com/dzhu/myo-raw`.