UNIVERSITÄT ZU LÜBECK

From the Institute of Robotics and Cognitive Systems
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Floris Ernst

Cutting-Edge Precision: Tissue Differentiation with an Intelligent Ultrasonic Aspirator

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
Niclas Erben
from Henstedt-Ulzburg

Lübeck, 2025

First referee: Prof. Dr. rer. nat. habil. Floris Ernst

Second referee: Prof. Dr.-Ing. Nele Rußwinkel

Date of oral examination: 06.05.2025

Approved for printing. Lübeck, 20.05.2025

Abstract

Medical professionals in neurosurgery face the critical challenge of accurately identifying tissue types during surgery. For instance, this involves determining where healthy tissue ends and tumor tissue begins, as well as adjusting equipment settings for optimal tumor resection. Surgeons must balance the complete removal of tumorous tissue while minimizing damage to healthy tissue, as damage to healthy tissue can have severe consequences, and any remaining tumorous tissue can lead to recurrence. The surgeons often rely on haptic sensations obtained by palpating the tissue to make these decisions, as the mechanical properties of the tissue can provide information about its type. However, this method is subjective and can be influenced by the surgeon's experience, stress, and fatigue. One tool that has been in use for tumor removal in neurosurgical operating rooms since the 1980s is the ultrasonic aspirator. As it exhibits similar technical characteristics to those of a tactile sensor, this handheld tool could potentially provide tactile information about the tissue type based on mechanical properties during tissue resection, without interrupting the surgical workflow.

This thesis investigates the feasibility of enhancing the ultrasonic aspirator with intelligence for tissue differentiation based on stiffness. The novel intelligent system consists of two modular components: contact detection and tissue differentiation, combined in a hierarchical manner. Both modules are data-driven, utilizing machine learning algorithms, and are evaluated from controlled experiments to real-world scenarios. To this end, various data sets were collected using artificial tissue models under different conditions, as well as in vivo human brain tissue. The contact detection module, utilizing a recurrent neural network architecture enhanced with advanced recalibration techniques and ensemble inference, demonstrates high performance. Key optimizations were implemented to minimize the missed detection rate of initial contacts, maintaining it at a very low level, which is essential for accurate subsequent tissue differentiation. The tissue differentiation module achieves metrics on par with or exceeding human capabilities, demonstrating the feasibility of using an ultrasonic aspirator for tissue differentiation. This differentiation between three distinct stiffness classes was achieved using a hierarchical classification approach with classic machine learning algorithms and a feature set designed to capture characteristics around the contact point. Additionally, uncertainties were quantified and communicated through the use of an ensemble of classifiers. Overall, the findings of this dissertation provide a basis for utilizing the ultrasonic aspirator as a tool for tissue differentiation and illustrate the potential of the intelligent system to assist surgeons in decision-making and other potential subsequent tasks during surgical procedures in the future.

Kurzfassung

Medizinische Fachkräfte in der Neurochirurgie stehen vor der Herausforderung, während der Operation die verschiedenen Gewebetypen präzise zu identifizieren. Dazu gehört unter anderem die Bestimmung, wo gesundes Gewebe endet und Tumorgewebe beginnt, sowie die Anpassung von Geräteeinstellungen für eine optimale Tumorresektion. Chirurg:innen müssen dabei ein Gleichgewicht zwischen der vollständigen Entfernung des Tumorgewebes und der Minimierung der Schädigung von gesundem Gewebe finden, da Schäden an gesundem Gewebe schwerwiegende Folgen haben können und verbleibendes Tumorgewebe ein Wiederauftreten des Tumors verursachen kann. Bei diesen Entscheidungen verlassen sich Chirurg:innen häufig auf haptische Wahrnehmungen, die sie durch das Abtasten des Gewebes erhalten, da die mechanischen Eigenschaften des Gewebes Aufschluss über dessen Art geben können. Diese Methode ist jedoch subjektiv und kann durch Erfahrung, Stress und Ermüdung der Behandelnden beeinflusst werden. Ein seit den 1980er Jahren in neurochirurgischen Operationssälen zur Tumorentfernung eingesetztes Instrument ist der Ultraschall-Aspirator. Da er ähnliche technische Charakteristiken eines taktilen Sensors aufweist, könnte dieses handgeführte Gerät während der Geweberesektion möglicherweise taktile Informationen über den Gewebetyp auf der Grundlage mechanischer Eigenschaften liefern, ohne dabei den chirurgischen Arbeitsablauf zu unterbrechen.

Diese Arbeit untersucht, ob es möglich ist, den Ultraschall-Aspirator mit einem intelligenten System zur Gewebedifferenzierung auf der Grundlage der Festigkeit zu erweitern. Dieses neuartige intelligente System besteht aus zwei modularen Komponenten: Kontakterkennung und Gewebedifferenzierung, die hierarchisch verbunden werden. Beide Module sind datengesteuert, nutzen Algorithmen des maschinellen Lernens und werden von kontrollierten Experimenten bis hin zu realen Szenarien evaluiert. Zu diesem Zweck wurden verschiedene Datensätze mit künstlichen Gewebemodellen unter verschiedenen Bedingungen sowie mit menschlichem in vivo Gehirngewebe gesammelt. Das Modul zur Erkennung von Kontakten, die eine rekurrente neuronale Netzwerkarchitektur nutzt, welches mit fortschrittlichen Rekalibrierungstechniken und Ensemble-Inferenz verbessert wurde, zeigt eine hohe Performance. Wichtige Optimierungen wurden vorgenommen, um die Rate der nicht erkannten Erstkontakte zu minimieren und auf einem sehr niedrigen Niveau zu halten, was für eine genaue nachfolgende Gewebedifferenzierung unerlässlich ist. Das Gewebedifferenzierungsmodul erreicht Metriken, die den menschlichen Fähigkeiten entsprechen oder diese sogar übertreffen, was die Machbarkeit der Verwendung eines Ultraschall-Aspirators für die Gewebedifferenzierung verdeutlicht. Diese Unterscheidung zwischen drei verschiedenen Festigkeitsklassen wurde mit Hilfe eines hierarchischen Klassifizierungsansatzes mit klassischen Algorithmen des maschinellen Lernens und einer Reihe von Merkmalen erreicht, die zur Erfassung von Eigenschaften rund um den Kontaktpunkt verwendet wurden. Zusätzlich wurden Unsicherheiten quantifiziert und durch den Einsatz eines

Ensembles von Klassifikatoren kommuniziert. Insgesamt bilden die Ergebnisse dieser Arbeit die Grundlage für den Einsatz des Ultraschall-Aspirators als Werkzeug zur Gewebedifferenzierung und zeigen das Potenzial des intelligenten Systems auf, Chirurg:innen zukünftig bei der Entscheidungsfindung und anderen potenziellen Aufgaben während chirurgischer Eingriffe zu unterstützen.

Contents

CHAPTER 1

Introduction

Neurosurgery is one of the most demanding fields in medicine, requiring exceptional precision, skill, and experience. Operating on the brain, one of the body's most delicate organs, means even the smallest mistake can lead to severe consequences. This is especially apparent in the removal of brain tumors, where surgeons have to balance the act of removing the tumor completely while at the same time damaging no surrounding healthy tissue. If too little of the tumor is removed, it can lead to recurrence, while removing healthy tissue can lead to neurological deficits. Moreover, the survival rate of the patient is directly correlated with the amount of tumor tissue that is removed [160, 166]. This is further aggravated by the fact that it can be difficult for surgeons to visually identify where the brain tumor ends and where the healthy tissue begins – an example of this is shown in Figure 1.1. The necessity of real-time assessment during

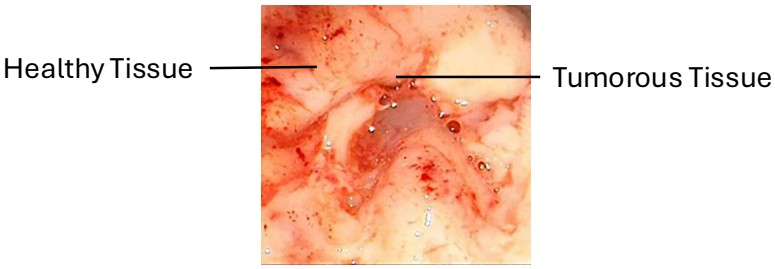


Figure 1.1: Example of tumorous tissue in the brain. The tumor is not clearly distinguishable from the healthy tissue in white light conditions.[1]

surgery, in a complex environment with such ultra-soft tissue, makes this even more difficult. To add to this, the removal of the tumor must be conducted in a time-efficient manner to minimize the patient's anesthesia exposure and prevent surgeon fatigue. Simultaneously, the procedure must be executed with precision to avoid inadvertent

1 Modified, Image by Barbara Kiesel et al., '5-ALA in Suspected Low-Grade Gliomas: Current Role, Limitations, and New Approaches', Frontiers in Oncology, July 30, 2021 [94]. Licensed under CC BY 4.0. (https://doi.org/10.3389/fonc.2021.699301)

damage to surrounding tissue. Among the various tools used in neurosurgery for intraoperative tissue differentiation – each with its own advantages and disadvantages – surgeons often depend on the haptic feedback from the tissue they are operating on. In other words, they rely on the sensations they feel when palpating the tissue. Such haptic feedback is highly subjective and experience-dependent, and can be influenced by various factors such as fatigue and stress.

A common tool used in neurosurgery is the ultrasonic aspirator. This handheld device uses ultrasonic vibrations along with aspiration and irrigation to break down and remove tissue. The speed with which the resection is done can be controlled by the surgeon or medical staff through the settings of the device. However, choosing optimal settings for the tissue at hand is user-dependent and subjective, and impacts the workflow of the surgeon. Since the working principle of the ultrasonic aspirator is similar to that of a tactile sensor, the concept of an intelligent ultrasonic aspirator emerges. Such a device could hold the potential to identify different types of tissue and act accordingly, for example by adjusting the settings automatically or by providing feedback to the surgeon about the tissue type. Instead of focusing on a direct differentiation of healthy and tumorous tissue, the intelligent component should be able to differentiate between tissues in their mechanical properties such as stiffness. This approach allows for a more general use of the device, with a multitude of tasks that can be tackled subsequently. Combining the resection capabilities of an ultrasonic aspirator with an intelligent system for tissue differentiation represents a novel approach to neurosurgery. Integrating intelligent components with instruments is not uncommon; an example can be seen in tools used for hemostasis.[1] These instruments are designed for vessel sealing and cutting, where an intelligent system automatically adjusts the energy supplied to minimize unnecessary thermal damage and enhances surgical outcomes [19, 116, 173]. However, these instruments lack the ability to provide intelligent insights about tissue types and function primarily as advanced control systems.

1.1 Goal of the Thesis

The aim of this thesis is to investigate the feasibility of making an ultrasonic aspirator intelligent. For this purpose, the problem is modeled as modular components, which are individually investigated and combined in the following flow chart, as shown in Figure 1.2. The basis for the intelligent ultrasonic aspirator forms the electrical data provided by its accompanying generator. To facilitate an intelligent system, this thesis concentrates on two modules: detecting contact between the ultrasonic aspirator and tissue, and differentiating between various tissue types. They are structured hierarchi-

1 THUNDERBEAT with Intelligent Tissue Monitoring, Olympus Medical Systems Corp., Tokyo, Japan
HARMONIC ACE+ with Adaptive Tissue Technology, Ethicon, Cincinnati, OH, USA
LigaSure, Medtronic, Minneapolis, MN, USA

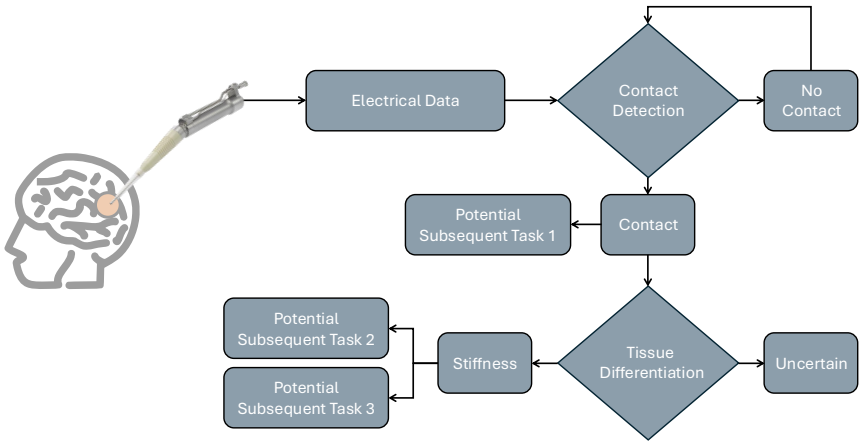


Figure 1.2: Flow chart of an intelligent ultrasonic aspirator.

cally, with contact detection being the basis for tissue differentiation, which can only proceed if contact detection is successful. Tissue differentiation then provides information about the stiffness of the tissue, or, if not enough information is available, reports its uncertainty accordingly. Each of these modules can be used to enable subsequent tasks. Thus, the primary research questions addressed in this thesis are:

I) Is contact detection feasible with an ultrasonic aspirator?

II) Is it feasible to accurately distinguish between different tissue types with an ultrasonic aspirator?

1.2 Structure of the Thesis

The remainder of this thesis is structured as follows. A medical and technical background is given in chapter 2 to provide the reader with a better understanding of the basics in the field of brain tissue and the ultrasonic aspirator. In the following chapter 3, a basic introduction to the algorithms and methods used in this thesis is provided. An overview of the data acquisition process and the data used throughout the following chapters is given in chapter 4. The investigation into the potential of the ultrasonic aspirator to detect contact with tissue is conducted in chapter 5. Subsequently, chapter 6 explores the feasibility of using the ultrasonic aspirator to differentiate between various tissue types. Finally, a conclusion is drawn in chapter 7, summarizing the results along with their limitations and providing an outlook on future work.

Each chapter is designed to be as self-contained as possible, ensuring that readers can follow the progression of ideas independently. Notably, chapter 5 and chapter 6 share a similar structure: they begin with a review of related work, followed by a series

of investigations addressing the problem under increasing difficulty from controlled to realistic conditions. Results are discussed comprehensively, culminating in a summary at the end of each chapter. In chapter 6, additional experiments are presented to evaluate the influence of external forces, insights from clinical data, and a comparison of the ultrasonic aspirator's performance in tissue differentiation against human capabilities.

CHAPTER 2

Medical and Technical Background

This chapter provides a concise overview of the fundamental knowledge required in the fields of the human brain, its anatomy, tumors, and its tissue properties. Furthermore, an introduction to the current state of the art in intraoperative tissue differentiation is given. Additionally, it offers a detailed description of the ultrasonic aspirator, encompassing both technical and operational aspects of the device.

2.1 Human Brain

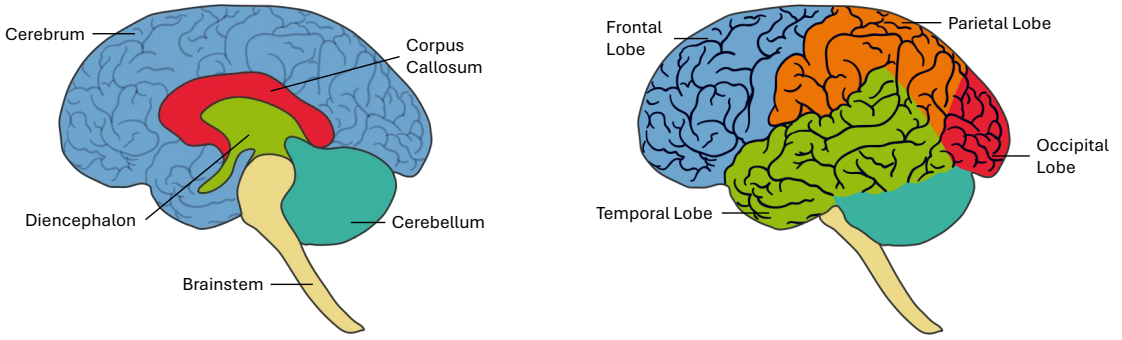
The human brain is the central organ of the human nervous system and is responsible for the control of most bodily functions, including thoughts, memory, emotions, and movements. It is therefore a complex and fundamental organ of the human body. Consequently, damage to the brain can have severe consequences, ranging from mild cognitive impairment to severe disabilities or even death. The intricacies of the human brain are vast and multifaceted, exceeding the scope of this work to delve comprehensively into its functionalities and diverse anatomical regions. Therefore, only an overview of the brain's complexities is provided here. For those seeking a deeper understanding, reference to clinical literature is recommended.

2.1.1 Basic Anatomy [1]

The human brain is situated in the cranial cavity, resting on the skull's base. It is safeguarded by the meninges, a three-layered membrane system. The outermost layer is the dura mater, followed by the arachnoid mater, and the innermost layer, the pia mater, which is directly adhered to the brain. Between the arachnoid mater and pia mater lies a layer of cerebrospinal fluid that cushions the brain. Together with the spinal cord, the brain forms the central nervous system (CNS) of the human body. Neurons are the basic building blocks of the brain and are responsible for transmitting information. Over 100 billion neurons are interconnected in the human brain, forming a complex network. Additionally, the brain consists of glial cells, which support the neurons in their metabolic and structural functions and make up ten to 50 times the number of

1 This subsection is based on the books of Bösel [14] and Kandel et al. [87]

neurons in the brain [182]. The brain can be divided into five main parts: the cerebrum, the cerebellum, the diencephalon, the corpus callosum, and the brainstem. A schematic overview of the human brain is shown in Figure 2.1(a) which highlights the basic structures of the brain in a sagittal view. Each part serves different functions, such as the cerebrum being responsible for higher brain functions, the cerebellum for motor control, the diencephalon for sensory processing, the corpus callosum for connecting the two hemispheres, and the brainstem for basic life functions such as breathing and heart rate regulation.



(a) Basic structures of the human brain. **(b)** The four lobes of the human brain.

Figure 2.1: Schematic overview of the human brain and its lobes.[1]

As can be seen in Figure 2.1(a), the cerebrum is the largest part of the brain, making up about 82 % of the brain's weight [74]. It is divided into two hemispheres, the left and right, connected by the corpus callosum. Gray and white matter are two types of tissue that make up the CNS. The gray matter is located on the surface of the cerebrum, while the white matter is located beneath the gray matter in deeper regions. In general, each hemisphere of the brain contains four distinguishable lobes: the frontal lobe, the parietal lobe, the temporal lobe, and the occipital lobe, as shown in Figure 2.1(b). Each lobe has specific functions. The frontal lobe is associated with cognitive tasks such as reasoning, planning, and problem-solving. The parietal lobe processes sensory information including touch, temperature, and pain. The temporal lobe handles auditory information and plays a role in memory and emotion. The occipital lobe is primarily responsible for processing visual information. However, the functions of these lobes are not strictly isolated; they interact with each other in complex ways.

1 Modified, Image by Cancer Research UK, licensed under CC BY-SA 4.0 (https://commons.wikimedia.org/w/index.php?curid=34334117)

2.1.2 Tumors of the Brain

Tumors of the brain are abnormal proliferations of cells. This abnormal growth is caused by the uncontrolled division of cells and inactivation or malfunction of the mechanisms that regulate cell growth and death. In the CNS, more than 150 different types of tumors can occur, which can be classified based on characteristics such as their location, histology and growth rate [110]. The World Health Organization (WHO) classifies CNS tumors into four grades which can be characterized by their growth rate, nature, and cellular activity and more as shown in Table 2.1. Generally speaking, the higher the

Table 2.1: Characteristics of CNS tumors by grade [109, 111]

Grade	Growth Rate	Nature	Cellular Activity
1	Slow	Benign	Low cellularity, non-infiltrative
2	Relatively slow	Intermediate	Moderately increased cellularity, slight infiltrative
3	Faster	Malignant	High cellularity, infiltrative, moderate to severe nuclear atypia
4	Very fast	Highly malignant	Very high cellularity, widely infiltrative, severe nuclear atypia, often with necrosis and microvascular proliferation

grade, the more aggressive the tumor is and the worse the prognosis and the more difficult the treatment [114]. Likewise, the infiltrative nature of the tumor becomes more pronounced with higher grades, as well as their increased cell density (cellularity) and abnormalities in the cell nuclei (nuclear atypia). Characterizing the tumor in terms of benign or malignant is important for the treatment and prognosis of the patient. Benign tumors are usually slow-growing and not invasive, meaning they do not spread to other parts of the body. Despite its non-invasive behavior, a benign tumor can cause damage and impairments in the brain by compressing surrounding tissues [177]. Malignant tumors, on the other hand, are fast-growing and invasive, meaning they can spread to other parts of the body. Causes of brain tumors can be genetic, lifestyle or due to environmental reasons such as exposure to ionizing radiation [115]. Brain tumors can be categorized based on their origin. Primary tumors start in the brain itself, whereas secondary tumors begin in other parts of the body and spread to the brain. Gliomas, which arise from glial cells, are the most common primary brain tumors, accounting for approximately 50 % of all primary brain tumors [130]. Meningiomas, which develop from the meninges, are the second most common, comprising about 20 % of primary brain tumors, of which 92 % are benign [130]. Gliomas can be further classified into different types, such as glioblastoma multiforme, which is the most aggressive and malignant type of glioma with a median survival of 14 months [118]. The most common

tumors are part of secondary tumors, which are metastases from other parts of the body. Almost half of all brain tumors are metastases, with the most common primary sites being the lung, breast, and skin [49]. Metastases are treated as malignant tumors as they are originating from different parts of the body. They reach the brain through the bloodstream or the lymphatic system and can cause severe damage to the brain tissue.

2.1.3 Tissue Properties of the Brain

The brain is the softest and most delicate organ in the human body. This can be seen in Figure 2.2. The only comparable material in terms of softness is adipose tissue which shows a similar stiffness to the brain. Still, the brain is even softer than adipose tissue. This makes brain tissue unique and challenging to work with, since it could be characterized as more fluid-like than solid-like and therefore be refered to as ultra-soft tissue.

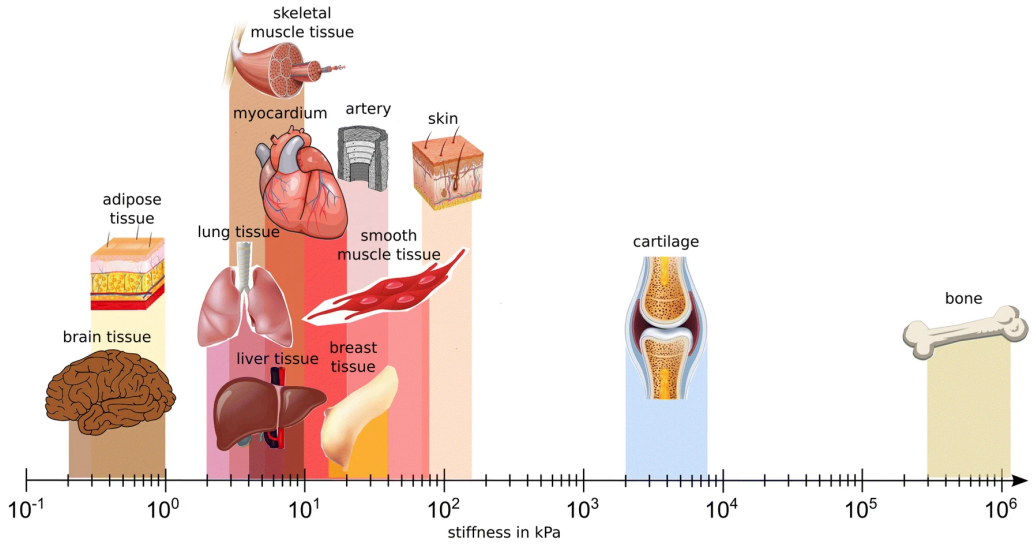


Figure 2.2: Stiffness levels of human body materials, with brain tissue being the softest.[1]

This is caused by the biphase structure of the brain, which consists of a solid phase and a liquid phase. The water content makes up about 80 % with 0.83 g/ml in gray matter and 0.71 g/ml in white matter [181, 56]. Moreover, deformations of the brain can already be observed just by its own weight in certain geometries [103], making measurements of tissue properties challenging [137]. Brain tissue can be characterized as a highly nonlinear, viscoelastic material [9]. Viscoelasticity describes the material's behavior as a combination of viscous and elastic properties, while the nonlinear behavior indicates that the material's response is not proportional to the applied load.

1 Image by Silvia Budday et al., 'Fifty Shades of Brain: A Review on the Mechanical Testing and Modeling of Brain Tissue', Archives of Computational Methods in Engineering, Springer Nature, July 19, 2019 [25]. Licensed under CC BY 4.0 (https://doi.org/10.1007/s11831-019-09352-w)

Stiffness properties can be described by a variety of mechanical parameters, such as Young's modulus, shear modulus or Poisson's ratio which are acquired through mechanical testing. These mechanical testing methods are having different setups and focus on different properties of the material. For example, uniaxial tensile or compression tests are used to determine the Young's modulus, which describes the stress-strain relationship of the material. Shear tests are used to determine the shear modulus, which describes the material's resistance to shear deformation. A Possion's ratio describes the material's behavior under compression or tension and is defined as the ratio of the lateral strain to the axial strain. For isotropic materials, a relationship between the Young's modulus, the shear modulus, and Poisson's ratio can be established by

$$E = 2G(1 + \nu), \tag{2.1}$$

where E is the Young's modulus, G is the shear modulus, and ν is the Poisson's ratio.

Measuring the stiffness of brain tissue is challenging due to its ultra-soft nature and complexity of the material. The following described challenges and known difficulties are mainly based on the review of Budday et al. [25]. Traditionally, mechanical testing was developed for materials way stiffer than brain tissue, which lead to problems in the discrepancy in results between early and recent studies [61]. Moreover, brain tissue proves to be fragile and will tear above certain thresholds under high strains for white matter material [56], subsequently impacting the results of the measurements. The aforementioned biphase structure of the brain tissue also complicates the measurements, as the incompressible liquid can flow out of the tissue during unconfined compression tests introducing time-dependent effects [137]. Thus, preconditioning of tissue and time for fluid recovery time between measurements are found to effect the results [21]. Furthermore, preconditioning can cause a softening of brain tissue after multiple cycles of loading and unloading [62]. Nevertheless, recovery of the tissue stiffness after preconditioning is observed [24]. The local functional dependencies of the brain are associated with varying cell structure, which in turn results in spatially dependent mechanical tissue properties, namely heterogeneous tissue [180]. Moreover, brain tissue becomes stiffer not only with increasing strain but also with increasing strain rate, i.e. the speed of deformation [24, 113, 136, 45], most likely related to fluid flow dependences which are more pronounced in white matter than in gray matter [20, 21, 23]. One of the highest difficulties in measuring brain tissue stiffness is caused by the limitation to the medium in its natural state [101]. Thus, only giving reduced numbers of studies with in vivo measurements [46, 52, 54, 56, 61, 84, 132, 158]. In vivo measurements are limited due to testing setup conditions requiring specific material geometries, which are not given in the brain. Additionally, multiple loading conditions in traditional setups are required which can cause damage to the tissue in healthy individuals. Deformation differences are to be expected between in vivo and ex vivo measurements, as the brain is not in its natural state in ex vivo measurements due to the absence of surrounding

tissue and skull. These differences remain an open question. All together, it is impossible to determine a single stiffness value for brain tissue, whether for gray matter, white matter or tumorous tissue. This is because testing conditions significantly influence the results, each measuring different aspects of stiffness. However, each test still provides comparable results within its own parameters. Generally, it is assumed that healthy brain tissue exhibits a stiffness in the order of 1000 Pa in terms of shear modulus [20].

2.1.4 Intraoperative Tissue Differentiation

Intraoperative tissue differentiation is the process of distinguishing between different types of tissue during surgery. In neurosurgery, the challenge is to differentiate between healthy and pathological tissue, given the potential for severe consequences for the patient if a wrong decision is made. This can either be removing healthy tissue, which can lead to severe impairments, or leaving pathological tissue behind, which can lead to tumor recurrence. In contrast, while other surgical disciplines also strive to minimize the margins around pathological tissue, they often have more flexibility, and the risks associated with these margins generally can be less critical compared to neurosurgery. Studies have shown that the extent of brain tumor resection is a factor in the prognosis of the patient [160, 166]. The greater the percentage of removed tumorous tissue, a significantly longer survival rate is observed, making complete resection the goal of neurosurgical operations. Thus, making accurate tissue differentiation crucial for the success of the operation which need to happen during the operation. Additionally, working in a highly complex environment is challenging, especially when dealing with the brain, one of the most delicate and important organs in the human body. Its ultra-soft nature adds to the difficulty.

A variety of methods can be used to enable surgeons to come to a decision on the tissue type at hand during surgery. Modern neurosurgery is supported by imaging techniques such as magnetic resonance imaging (MRI) and computed tomography (CT) scans which provide preoperative information on the tumor location and its relation to surrounding structures. This information is combined with neuronavigation systems which allow the surgeon to navigate through the brain during surgery. However, these methods lack the ability to provide real-time information on the tissue type, as the images are taken before the surgery and do not account for changes in the brain during the operation which inevitably happen as the brain shift takes place when the skull is opened [65]. A promising method for making the tumor more visible during surgery is the use of fluorescent agents which are injected into the patient. The agent accumulates in the tumorous tissue and can be visualized using a fluorescence microscope which is equipped with a light source that excites the agent in the required spectrum. Nevertheless, this method is not generally applicable to all tumor entities as they are not all equally sensitive to a contrastive agent [88, 157, 2, 129]. Other intraoperative imaging techniques exist, such as intraoperative MRI or 3D ultrasound [37, 143, 29, 140,

176, 144, 78]. Methods such as intraoperative MRI have not yet become part of routine clinical practice in neurosurgery. This is primarily because of their high cost and the complex infrastructure necessary to operate these systems. Likewise, intraoperative 3D ultrasound has faced challenges in becoming a widely adopted standard tool within the neurosurgical operating room. Despite its potential benefits, the technology has not yet established itself as a user-friendly component of the neurosurgical intraoperative workflow. This is largely due to the complexities involved in interpreting the images it produces, which can often be difficult for surgeons to accurately understand and utilize during procedures. Consequently, its integration into routine neurosurgical practice has been limited. Currently, newer methods are under investigation, including Raman spectroscopy, optical coherence tomography and intraoperative in vivo confocal laser endomicroscopy imaging [164, 162, 163, 76, 55, 141, 142, 1]. These advanced techniques often require significant additional effort in the workflow for surgeons and other medical professionals. As a result, they have not yet been established as part of the routine clinical practice. Therefore, surgeons still rely heavily on their experience and haptic feedback to distinguish between tumor and healthy tissue. This haptic feedback depends on the tissue properties, such as stiffness, which can differ between pathological and healthy tissue [20, 21, 22, 25, 81, 139, 26]. However, this method is highly subjective and can be influenced by various factors, including the surgeon's fatigue, stress, and level of experience. Consequently, different tactile sensors are actively being researched to provide objective information on tissue properties during surgery [170, 85, 165]. However, these methods usually require additional equipment and are not part of the standard neurosurgical workflow. Therefore, it is highly desirable to have an objective method for tissue differentiation that seamlessly integrates into the surgeon's workflow, such as an intelligent ultrasonic aspirator.

2.2 Ultrasonic Aspirator

Since the early 1980s, a commonly used neurosurgical instrument for tumor removal is the ultrasonic aspirator [18, 73]. It combines three functions of ultrasonic-based fragmentation, aspiration and irrigation in one device – all working simultaneously to ensure efficient and safe tumor removal. In contrast to diagnostic and therapeutic ultrasound in the medical area, the ultrasonic aspirator employs surgical ultrasound. Diagnostic ultrasound is used as an imaging technique to visualize the inside of the body. It uses a frequency range of 1 to 15 MHz and is non-invasive and free of ionizing radiation. Therapeutic ultrasound is used for therapeutic purposes, such as pain relief by heating the tissue in deep regions of the body or facilitating tissue regeneration processes [68]. It uses a frequency range of 0.7 to 1 MHz and is also non-invasive. Surgical ultrasound, on the other hand, can be used for tissue removal or kidney stone fragmentation and employs a lower frequency range of 20 to 40 kHz while maintaining high intensity levels [119]. All three types of medical ultrasound applications are based on the same physical

principle that creates the desired effect. If a deformation is applied to a piezoelectric crystal, an electric voltage is generated – called the piezoelectric effect [179]. Conversely, if an electric voltage is applied to a piezoelectric crystal, it deforms. If an alternating voltage is applied, the crystal thus will deform periodically, creating an oscillation.

2.2.1 Components and Tissue Removal Mechanism

The ultrasonic aspirator on a basic level consists of two main components: the handpiece and the high frequency generator. Both components are displayed in Figure 2.3 and are connected via three cables/tubes – facilitating the transmission of electrical signals, aspiration, and irrigation. A metal-based hollow tube, known as the sonotrode, is

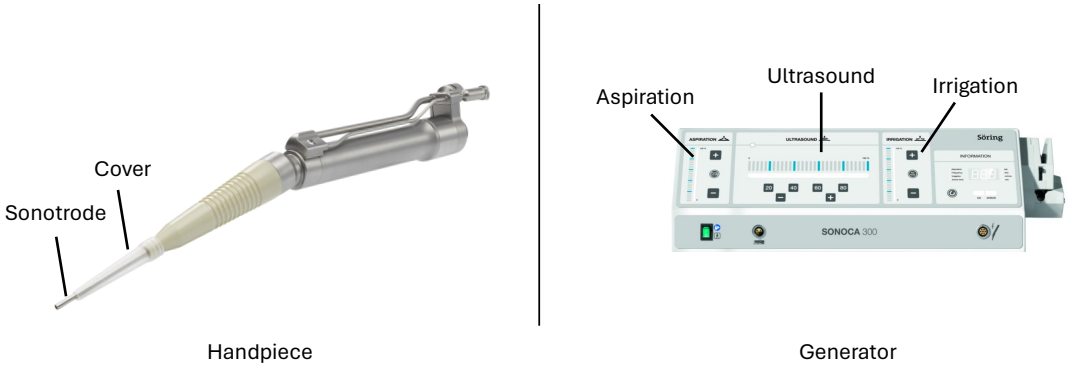


Figure 2.3: Components of the ultrasonic aspirator. The handpiece (left) and the generator (right) are connected via three cables/tubes. The handpiece holds the sonotrode at the end which contacts the tissue and is surrounded by a cone-shaped cover. The generators interface allows a user to control the ultrasonic oscillation of the sonotrode, the aspiration and irrigation independently.[1]

situated at the distal end of the handpiece. This sonotrode contacts the tissue and is crucial for conducting and amplifying the ultrasonic vibrations. The aforementioned functions all serve a distinct purpose in removal process and are visualized in Figure 2.4. The ultrasonic-based fragmentation uses the mechanical oscillation of the sonotrode to break down the tissue cells – similar to a jackhammer. Aspiration uses negative pressure to remove the fragmented tissue through the hollow center part of the sonotrode. In addition, increasing suction pressure can increase the stress on the tissue, which leads to better tissue removal – especially in soft tissue. Irrigation adds a continuous flow of saline solution to cool the sonotrode and ensure moistening of the tissue to facilitate efficient tissue fragmentation. The saline solution is transported through the cone-shaped cover of the sonotrode to the tissue, by that being able to cool the sonotode. Mainly, three effects are responsible for the tissue removal: cavitation, mechanical

1 Images with permission by Söring GmbH, Quickborn, Germany

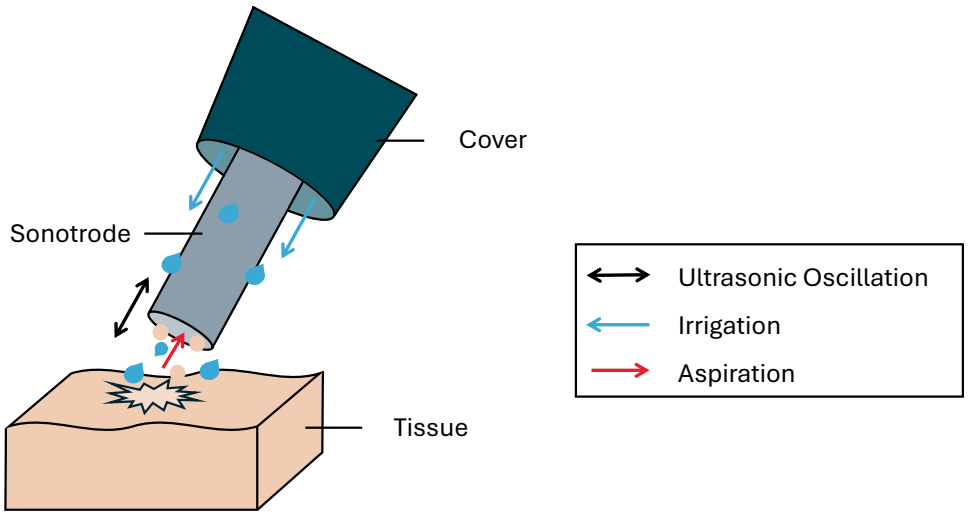


Figure 2.4: The three functions of the ultrasonic aspirator: ultrasonic oscillation of sonotrode, irrigation and aspiration. All applied simultaneously to ensure efficient and safe tissue removal.

stress and thermal effects [123]. However, the precise impact of these mechanisms on fragmentation remains uncertain and the combination of these mechanisms has yet to be definitively established [127, 35]. The extent of these effects may differ based on the structural integrity and the interaction between these effects. For instance, a higher amount of structural protein increases resistance, making the tissue more difficult to fragment [35, 98]. As a result, the fragmentation process is likely to be dependent on the properties of the tissue. Thus, the high-frequency generator enables the user to control the fragmentation and manage the three functions of the ultrasonic aspirator. The user can adjust the ultrasound in percentage, implicitly controlling the amplitude of the sonotrode, the aspiration pressure in mbar and the irrigation flow rate in ml/min. This provides the user with a multitude of potential configurations for the settings. A foot pedal is used to activate the ultrasonic aspirator by pressing it. Releasing the foot pedal deactivates the ultrasonic aspirator and stops the fragmentation process.

In everyday clinical practice, the ultrasonic aspirator is used as follows: the surgeon holds the handpiece while the generator remains outside the sterile area. The surgeon starts and stops the ablation process using a foot switch. If the tissue requires a setting adjustment, the surgeon cannot make this adjustment directly but must verbally instruct a medical staff member to do so. This takes place under the inherently stressful conditions of a neurosurgical tumor operation.

2.2.2 Operating Principle

As established, the ultrasonic aspirator is a multifunctional device that combines ultrasonic-based fragmentation, aspiration and irrigation. The high frequency generator – aside from controlling the irrigation and aspiration pumps separately – is responsible for generating the ultrasonic vibrations. The general principle is as follows and visualized in Figure 2.5: The generator produces an alternating voltage. This electrical output

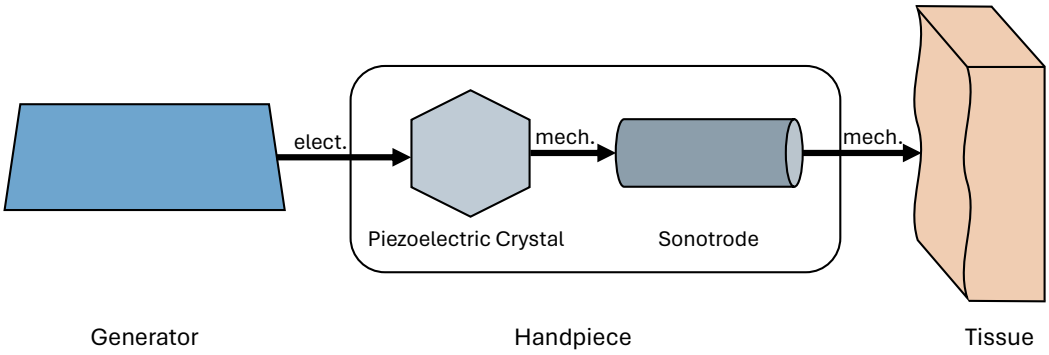


Figure 2.5: Operating principle of the ultrasonic aspirator. An electrical output is applied to the piezoelectric crystal which deforms periodically, creating a mechanical oscillation. Subsequently, the mechanical oscillation is amplified by the sonotrode and transmitted to the tissue.

is applied to the piezoelectric crystal located in the handpiece. The crystal deforms periodically, creating a mechanical oscillation in longitudinal direction. This mechanical oscillation is amplified by the sonotrode and transmitted to the tissue. Due to proprietaries of the specific workings of the high frequency generator, a detailed description of the generator's inner workings is not possible and outside the scope of this work. A variety of different electronic components such as direct and alternating current converters, amplifiers and resonant circuits, consisting of an inductor and a capacitor, are used to generate a specific frequency and amplitude. When the ultrasonic aspirator is started by pressing the foot pedal, a frequency search is performed to find the resonant frequency of the handpiece. For this, a frequency range in the area of 35 kHz is scanned until the resonant frequency is found. In order to guarantee the continued functionality of the system, it is controlled according to a specific electrical parameter. The resulting alterations in the electrical values permit the efficient ablation of tissue, even in the event of sudden loads caused by contact with the tissue.

A comparable principle is evident in the field of tactile sensors. Established in 1968, the ultrasonic contact impedance method employs a piezoelectric element to measure the hardness of a material [97]. Primarily designed for testing metals, a probe oscillates

at its resonant frequency. Upon contact with the material, the oscillation is dampened, allowing the hardness to be calculated from the shift in resonant frequency. Similarly, advances have been made in the field of tactile sensors for medical applications [125, 124, 106, 107, 170, 85, 165]. Unlike the ultrasonic aspirator, these sensors are specifically designed for tactile sensing tasks and therefore have a different setup and working principle, such as maintaining the sensor in resonance through a feedback loop.

CHAPTER 3

Algorithmic Background

This chapter outlines the algorithmic foundations utilized in this thesis, divided into two main sections. The first section covers data-driven prediction techniques, while the second section focuses on the prediction models applied throughout the thesis.

3.1 Data-Driven Prediction Techniques

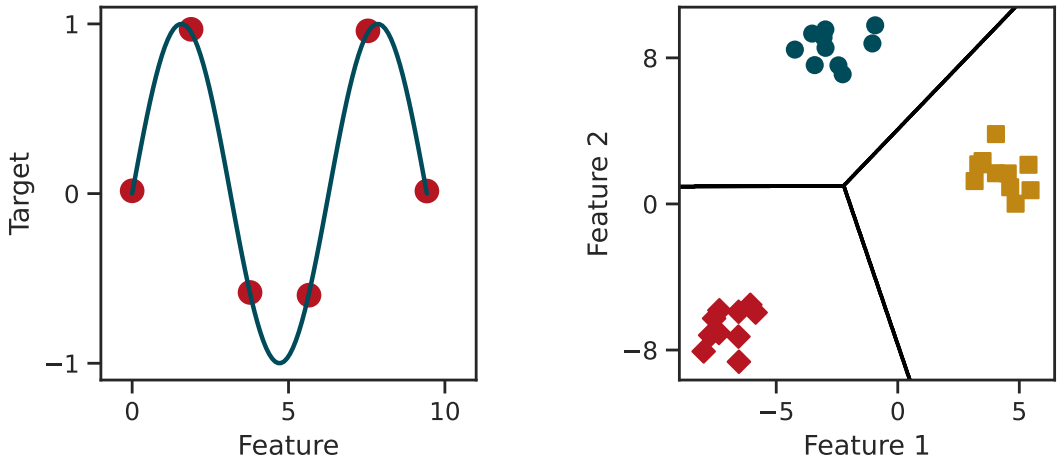
Predicting certain events or values is a common task in many fields. These predictions rely on collected data, hence the term data-driven predictions. In this context, the input data are referred to as independent variables and the output data are known as dependent variables. In machine learning, these are often called features and labels, targets, or classes. A prediction model maps input data to output data. The general form of a prediction model can be represented by the following equation:

$$\hat{y} = f(x) \tag{3.1}$$

where $\hat{y}$ is the predicted value, f is the prediction model, and x is the input data. The objective is to develop a prediction model that accurately maps the input data to the output data. This process is generally categorized under supervised learning in the field of machine learning. In supervised learning, the model is trained on labeled data, meaning the output data are known. Conversely, in unsupervised learning, the model is trained on unlabeled data, where the output data are unknown. Different types of predictions can be made for $\hat{y}$, which are outlined in the following paragraphs.

Regression Regression is a prediction technique that is used to estimate a continuous value based on input data. The goal is to determine a function that maps the input data to the output data, represented as $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Essentially, regression is a statistical method for fitting a mathematical function that best describes the relationship between the input and output data. Figure 3.1(a) illustrates this with six 1D data points and a fitted function. A real-world example of regression is predicting the current phase of a breathing cycle using data from tracked markers on the chest.

Classification Classification is a prediction technique that is used to predict a discrete value based on input data like $f : \mathbb{R}^n \rightarrow \{1, \dots, k\}$. k is the number of classes or categories that the output data can take. Decision boundaries are used to separate the different classes. This is shown in Figure 3.1(b) where the data points are separated into three classes by decision boundaries. In this example, the input data are two-dimensional and the output data are one of the three classes, seperated by lines. In multidimensional data, the decision boundaries are hyperplanes. Real-world applications of classification can be found in automatic image recognition where the input data is an image and the output data is the class of the image, e.g. cat, dog or car. Such a classification problem is called flat classification, since direct inference is made from the input data to the output data without any intermediate steps in the sense of a hierarchy. In most fields, this is the default way of classification.



(a) Example of a regression problem.

(b) Example of a classification problem.

Figure 3.1: Examples of regression and classification problems. The left figure shows a regression problem where the goal is to predict a continuous value. Red dots represent the input data and the blue line represents the prediction model. The right figure shows a classification problem where the goal is to predict a discrete value. Different colors represent different classes and the black lines represent the decision boundary.

Hierarchical Classification[1] Hierarchical classification is a special form of classification. In hierarchical classification, the classes are organized in a hierarchy, due to the taxonomic structure of the classes. This means that the classes are not on the same level, but are organized in some kind of graph or tree structure where the root node is the most general class and the leaf nodes are the most specific classes. The goal of hierarchical

1 This paragraph is based on the review paper of Silla et al. [159]

classification is to predict the most specific class for a given input data. Between the nodes, an *is-a* relationship is defined which means that a child node is a subclass of its parent node while being asymmetric, anti-reflexive and transitive. This means that an *is-a* relationship is irreversible, a child node cannot be related to itself, and if a child node is a subclass of a parent node, then the child node is also a subclass of the grandparent node. Regarding the previous example of automatic image recognition, a hierarchical classification could be used to first classify whether the image shows an animal or a car and then classify the animal into a cat or a dog. Thus, getting more specific with each step. An example of how a hierarchical classification structure could look like is shown in Figure 3.2. The first node is the root node which is the beginning of the hierarchy, followed by the classes A and B. Class A gets further divided into classes A1 and A2, while class B gets further divided into classes B1 and B2. Those four classes represent the leaf nodes of the hierarchy.

Different approaches exist to solve hierarchical classification problems. On a general scope, they can be seen as global or local approaches. Global approaches are based on the entire hierarchy and try to optimize the classification of all classes at once – including internal nodes as opposed to a flat classification which only considers leaf nodes. The difference between a flat classification and a global approach is shown in Figure 3.2. Such global approaches are commonly known as global classifier or big-bang

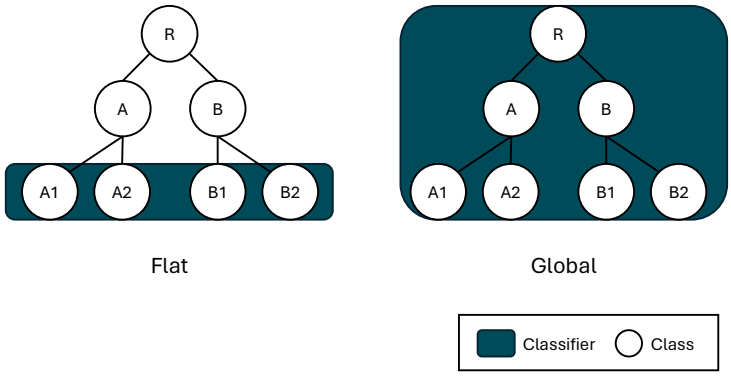


Figure 3.2: Flat classification and global approach for hierarchical classification. The left shows the flat classification where only leaf nodes are considered for classification. The right shows the global approach where all nodes are considered for classification. White circles represent the classes in the hierarchy, blue boxes represent classifiers predicting the enclosed classes.

approaches. Contrary, local approaches are based on the local structure of the hierarchy and try to optimize the classification of a single node or subset of nodes. They are further divided into local classifier per node, per level and per parent approaches. The local classifier per node approach uses a binary classifier for each node in the hierarchy, except

for the root node. The local classifier per level approach uses a classifier for each level in the hierarchy. And the local classifier per parent approach uses a classifier for each parent node in the hierarchy. A representation for each of these local approaches are shown in Figure 3.3. Training the necessary classifiers for each approach can differ, given

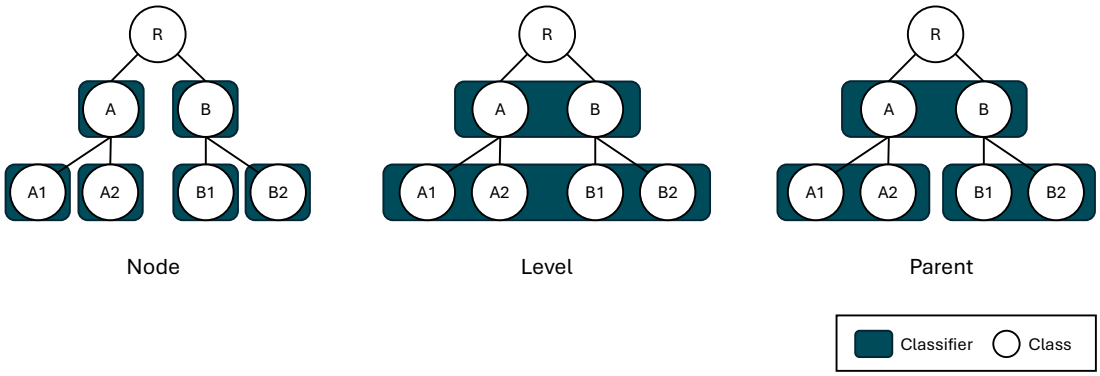


Figure 3.3: Local classifier per node, per level and per parent approaches for hierarchical classification. White circles represent the classes in the hierarchy, blue boxes represent classifiers predicting the enclosed classes.

the necessity of declaring data as positive or negative examples. Multiple policies exist such as exclusive [50], less exclusive [50], less inclusive [50, 53], inclusive [50], siblings [53, 28] and exclusive siblings [28]. Details on these policies are found in the respective literature and are not further discussed here.

Testing hierarchical classification is done by traversing the hierarchy from the root node to the leaf nodes, commonly known as a top-down approach [99]. The choice of which classes to predict for each level is based on the prediction of the parent node. This is done until a leaf node is reached. This is referred to as mandatory leaf node prediction (MLNP) [57]. Some approaches need specific strategies to handle such a traversal, e.g. the local classifier per level approach could predict a class not belonging to a previous predicted parent class. For example, starting from the root node, the classifier predicts class A, but in the following level, the outcome is class B1. In this specific case, output restrictions can be used to prevent such misclassifications. Similarly, local classifier per parent and per node only consider the classifiers of the parent or the node which is the child of the parent. All methods share that a wrong prediction in an upper level is propagated to the lower levels.

To circumvent this problem, non-mandatory leaf node prediction (NMLNP) is used [169]. This approach allows the hierarchical classification to stop at any level and not necessarily predict a leaf node. This can be done by only allowing transitions to child nodes if the classifier is confident enough. Accordingly, this is at the expense of outputting a less specific class.

It should be noted that hierarchical classification is not a classification approach itself, but a way to structure and traverse the output of a tree-like network of classifiers. Thus, the choice of the classifier is independent of the hierarchical classification approach and can be any classifier that is suitable for the problem at hand. Specifically, this allows to use different classifiers for different nodes in the hierarchy which can focus on the specific characteristics of the classes and leverage different features [156].

3.2 Prediction Models

Finding a mapping between input data x and output data $\hat{y}$ is done by a prediction model f as shown in Equation 3.1. There are various prediction models that can be used to solve this task which are outlined in the following. The level of detail is kept to a minimum to provide a general overview of the models and its characteristics. For more detailed information, the respective literature should be consulted.

3.2.1 Classical Machine Learning

Classical machine learning refers to the traditional machine learning models that are used to solve prediction tasks. They often involve manually crafted features that are used to train the model and require small amounts of data. They can provide interpretable results, computational efficiency and effectiveness.

Random Forest [17] Random Forest (RF) is an ensemble learning method that is used for classification and regression tasks. It is based on decision trees where multiple trees are trained on different subsets of the data, hence the description as an ensemble method. The final prediction is made by averaging the predictions of the individual trees. Random Forest is known for its robustness against overfitting and its ability to handle high-dimensional data.

AdaBoost [58] Similar to RF, AdaBoost is an ensemble learning method that is used for classification and regression tasks. The key idea of AdaBoost is to train multiple weak learners on the data where each learner is trained on the weighted misclassified data of the previous learner. Thus focusing on the difficult examples and minimizing an exponential loss. Weak learners are models that are only slightly better than random guessing, like decision stumps. The final prediction is made by the weighted sum of the individual learners, creating a strong learner.

Gradient Boosting [59] Gradient boosting and AdaBoost are similar in the sense that they are both ensemble learning methods that learn from the mistakes of the previous learners. The difference is that, unlike AdaBoost, each learner tries to minimize a residual error by using gradient descent on specific loss functions. Generally, gradient boosting

uses more complex weak learners. This method is known for its high flexibility and ability to handle complex data. Popular implementations of gradient boosting are Extreme Gradient Boosting (XGBoost) [30] and Light Gradient Boosting Machine (LightGBM) [90], which are optimized for speed and performance.

Support Vector Machine [38] Support vector machines (SVMs) are well-known models that are used for classification and regression tasks. SVMs are based on the idea of finding the hyperplane that best separates the classes in the data for classification tasks. The hyperplane is found by maximizing the margin between the classes. This is based on the support vectors which are the data points that are closest to the hyperplane. Due to the kernel trick, SVMs can handle non-linear data by transforming the data into a higher-dimensional space.

k-Nearest Neighbors [40] k-nearest neighbors (kNN) is a simple and intuitive model. The prediction is made by finding the k nearest neighbors of the input data in the feature space and taking the majority vote of the neighbors. The choice of k is crucial for the performance of the model. Furthermore, the distance metric is important for the performance of the model and should be chosen wisely according to the data.

Naive Bayes [47] Naive Bayes is a probabilistic model that is based on Bayes' theorem. The naive assumption is that the features are independent of each other, which is often not the case in practice. However, this assumption simplifies the model and makes it computationally efficient and effective. It calculates the probability of the classes given the features and makes the prediction based on the highest probability.

Logistic Regression [41] Logistic regression, despite its name, is a model that is used for classification tasks. It is based on the logistic function (sigmoid function) that maps the input data to the output data. The logistic function is used to calculate the probability of the classes given the features. The prediction is made by setting a threshold on the probability. In its simplest form, it is a linear model that is used for binary classification tasks. However, it can be extended to multi-class classification tasks and non-linear data by using different techniques like one-vs-rest or kernel tricks.

Gaussian Process [138] Gaussian process (GP) is a probabilistic model that is mainly used for regression tasks. It is based on the assumption that the data are generated by a Gaussian process. The model is defined by a mean function and a covariance function. The prediction is made by calculating the posterior distribution of the data and the prior. This prior is defined by the mean and covariance function. Common choices for the covariance function include the squared exponential and the Matérn kernel. Choosing

the right kernel is crucial for the performance of the model. Advantages of GPs are that they provide uncertainty estimates and are computationally efficient for small datasets.

Linear Discriminant Analysis [60] Linear discriminant analysis (LDA) can be used for classification tasks as well as a dimensionality reduction techniques often employed in a preprocessing step. The goal of LDA is to find the linear combination of features that best separates the classes of the data. In such, it tries to maximize the between-class variance and minimize the within-class variance. An assumption of LDA is that the classes have the same covariance matrix. As the name suggests, LDA is a linear model and can be used for binary and multi-class classification tasks. LDA shares similarities with principal component analysis (PCA) but focuses on the separation of the classes rather than the variance of the data. Furthermore, PCA is limited as a dimensionality reduction technique since it does not consider the class labels.

Quadratic Discriminant Analysis [60] Quadratic discriminant analysis (QDA) is similar to LDA but allows for non-linear decision boundaries. The difference is that QDA does not assume that the covariance matrix of the classes is the same. Thus, it estimates the covariance matrix for each class separately. Accordingly, QDA is more flexible than LDA but can be computationally more expensive.

Change Point Detection [5] Change point detection (CPD) is a statistical method used to identify points in a time-series where the statistical properties of the data change. Thus, making this not a prediction model but rather a method to detect changes in the data. This method can be useful for detecting shifts in data patterns. CPD algorithms typically monitor data sequentially to detect changes in mean, variance or other statistical properties. Bayesian online CPD, a common approach, employs a probabilistic framework to continuously update the belief about the presence of change points as new data arrives [3]. A constant hazard function is often used as a prior, reflecting the expected rate of change points within the time-series.

Perceptron [146, 147] The perceptron represents a simple linear model that is mainly used for binary classification tasks. It is based on the idea of a single-layer neural network and consists of weights and a bias term as well as an activation function. The weights are updated based on the error of the prediction using the perceptron learning rule. The perceptron forms the basis of neural networks and can be extended to multi-layer neural networks to handle more complex data. The basic perceptron is defined as

$$f(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x} + b), \tag{3.2}$$

where **x** are the input data, **w** are the weights, b is the bias term and θ is the activation function.

3.2.2 Deep Learning [1]

The previous section outlined classical machine learning models that are used to solve prediction tasks. Despite the advantage of interpretability and low computational costs, classical machine learning models do require manual feature engineering. Deep learning models, on the other hand, are able to learn the features from the data itself which can lead to better performance, especially on complex data like images, signals or text. Generally, the term deep learning refers to neural networks that have multiple layers consisting of neurons. The learning process consists of updating the weights of the neurons based on the error of the prediction. Backpropagation is used to calculate the gradient of the error with respect to the weights [149]. This gradient is then used to update the weights of the model using optimization algorithms like stochastic gradient descent [149] or Adam [95] and its variants like RAdam [108]. Different non-linear activation functions are used to introduce non-linearity into the model, allowing it to learn complex patterns in the data. Most common activation functions are the sigmoid, tanh, and ReLU [120]. During the training process, the model is trained over multiple epochs where the entire dataset is passed through the model in batches. Different processes can be used to prevent memorization of the data, called overfitting, such as dropout [161], batch normalization [79], early stopping [133] or regularization techniques like L1 or L2 regularization [122]. The following outlines different deep learning models architectures that are used to solve different tasks.

Fully-Connected Neural Network [147] A fully-connected neural network (FCN) is the simplest form of a neural network, also known as a multi-layer perceptron. The network consists of an input layer, multiple hidden layers and an output layer. Each layer consists of neurons that are connected to the neurons of the previous and the next layer. Hence, the name fully-connected. A simple stacking of two layers is represented as

$$f(\mathbf{x}) = \theta(\mathbf{w}_2^T \theta(\mathbf{w}_1^T \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2), \tag{3.3}$$

where **x** is the input data, $\mathbf{w}_1$ and $\mathbf{w}_2$ are the weights of each layer, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the respective bias terms and θ is the activation function. FCNs can be used for regression and classification tasks, simply depending on the output layer and the loss function that is used. Due to the fully-connected nature, the number of parameters can grow rapidly with the number of neurons and layers, which can lead to high computational cost and overfitting [64]. Furthermore, FCNs are not well-suited for data with spatial or

1 This subsection is based on the book of Goodfellow et al. [66]

temporal structure like images or signals since structure information can be lost during the flattening of the data.

Convolutional Neural Network [105] The convolutional neural network (CNN) is a neural network that is specifically designed to handle spatial data like images. CNNs have gained popularity and commercial attention with the success in the ImageNet competition in 2012 [100]. However, CNNs were already introduced in the 1980s by LeCun et al. [105] and only recently gained popularity due to the increase in computational power and the availability of large datasets. The key difference between CNNs and fully-connected neural networks is the use of convolutional layers. These allow to use convolutional procedures instead of general matrix multiplications shown in Equation 3.3. A simple convolutional layer for a two-dimensional input data is defined as an input I to an output O with a kernel K as

$$O(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n), \tag{3.4}$$

where i and j are the indices of the output data, m and n are the indices of the kernel. The kernel is learned during the training process and is used to extract features from the input data. Using convolutions advantages the model by reducing the number of parameters and by preserving the spatial structure of the data. Furthermore, the weights of the kernel are shared across the input data which allows the model to learn translation-invariant features. CNNs often consist of multiple convolutional layers followed by pooling layers such as max-pooling or global average pooling to reduce the spatial dimensionality of the data. Depending on the architecture, CNNs can have multiple branches [6], skip connections [145], residual blocks [72] or attention mechanisms [178] to improve the performance of the model.

Recurrent Neural Network [149] The recurrent neural network (RNN) is a neural network that is specifically designed to handle temporal data like time-series or sequences. RNNs have the ability to remember past information and use it to make predictions. They were first introduced in the 1980s by Rumelhart et al. [149] and later improved by Hochreiter and Schmidhuber [75]. This is done by using recurrent connections that allow the model to pass information of different states s to the next state. This can be represented in its simplest form as

$$s_t = f(s_{t-1}, x_t), \tag{3.5}$$

where s_t is the state at time t, f is the model function, s_{t-1} is the state at time $t-1$ and x_t is the input data at time t. Different architectures of RNNs exist like the long short-term memory (LSTM) [75] and the gated recurrent unit (GRU) [33]. LSTMs and

GRUs are specifically designed to handle the vanishing gradient problem that occurs in simple RNNs. Those architectures have gates that control the flow of information and gradients through the network, allowing the model to learn long-term dependencies.

Temporal Convolutional Network [7] The temporal convolutional network (TCN) is a neural network architecture that is specifically designed to handle time-series data – similar to RNNs. However, TCNs are based on convolutional layers instead of recurrent connections. It leverages one-dimensional convolutions along the time axis to learn the temporal structure of the data. Additionally, dilated convolutions are employed that increase the receptive field of the model and allow the model to learn long-term dependencies. Due to only using causal convolutions, TCNs can be used for online prediction tasks, since the model only uses past information to make predictions. TCNs are known for their parallelism and computational efficiency compared to RNNs, which can be slower due to the sequential nature of the model during training.

Autoencoder [15] Autoencoders are neural network architectures that are used for unsupervised learning tasks. The goal of an autoencoder is to learn a compressed representation of the input data. This is done by training the model to reconstruct the input data at the output layer, while passing the data through a bottleneck layer. This bottleneck layer is the compressed representation of the input data, often referred to as the latent space. The autoencoder consists of an encoder e that maps the input data to the latent space and a decoder d that maps the latent space to the output data. This is represented as

$$x = d(e(x)), \tag{3.6}$$

where x is the input data. Hence, this representation learning motivates the model to learn the most important features of the data that are necessary for the reconstruction. It shares similarities with dimensionality reduction techniques like PCA or LDA, but is more flexible and can learn non-linear features. Since autoencoders are architectures, they can be used with both fully-connected or convolutional layers, depending on the data and problem at hand.

3.3 Evaluation Metrics

The performance of a prediction model can be evaluated using different metrics. The choice of the metric depends on the task at hand, e.g. regression or classification, and the characteristics of the data. For regression tasks, common metrics are the root mean squared error (RMSE), the mean absolute error (MAE), or the coefficient of determination

(R^2). These three metrics are defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{3.8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.9}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value of n test data points and $\bar{y}$ is the mean of the true values.

Classification tasks are evaluated using metrics such as accuracy (ACC), precision or positive predictive value (PPV), recall or true positive rate (TPR), and the F1-score (F1). The definitions of these metrics are given as

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.10}$$

$$\text{PPV} = \frac{TP}{TP + FP}, \tag{3.11}$$

$$\text{TPR} = \frac{TP}{TP + FN}, \tag{3.12}$$

$$\text{F1} = \frac{2 \cdot \text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}}, \tag{3.13}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. Those numbers can be derived from a confusion matrix that is created by comparing the true values with the predicted values as shown in Figure 3.4.

Additionally, the area under the receiver operating characteristic curve (AUROC) can be used to evaluate the performance of a classification model to find the best trade-off between true positive rate and false positive rate. Specifically in case of imbalanced data, weighted variants of the aforementioned metrics can be used to account for the class distribution, as well as the balanced accuracy (BA) which is the average of the sensitivity and specificity, which is defined as

$$\text{BA} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \tag{3.14}$$

Since evaluations are mostly designed for flat classification, additional metrics can be utilized to evaluate the performance of hierarchical classification approaches: the hierarchical positive predictive value (hPPV), the hierarchical true-positive rate (hTPR)

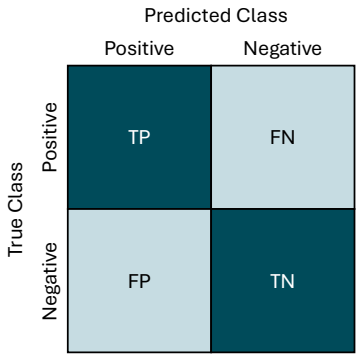


Figure 3.4: Example of a confusion matrix. The rows represent the true classes and the columns represent the predicted classes. The diagonal elements represent the correct predictions, while the off-diagonal elements represent the incorrect predictions.

and the hierarchical F1 (hF1) [96]. The metrics are defined as follows:

$$\text{hPPV} = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{P}_i|}, \tag{3.15}$$

$$\text{hTPR} = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{T}_i|}, \tag{3.16}$$

$$\text{hF1} = \frac{2 \cdot \text{hPPV} \cdot \text{hTPR}}{\text{hPPV} + \text{hTPR}}, \tag{3.17}$$

where $\hat{P}_i$ is the set of all predicted classes along the path from the root to the most specific predicted class for test data i. Similarly, $\hat{T}_i$ is the set of all true classes along the path from the root to the most specific true class for test data i.

CHAPTER 4

Data

Collecting data is a crucial part of any data-driven approach. However, acquiring data in the context of brain tissue presents a challenge due to the invasive nature of neurosurgery, which limits the availability of such data. Additionally, data collected under surgery conditions inherently contains a high number of unknowns, such as the exact type of tissue, surgeons' behavior and other environmental factors. Therefore, to overcome these limitations and to be able to collect a sufficient amount of data, acquisition is focused on artificial tissue models in a laboratory environment. Nevertheless, to be able to evaluate the transferability to the clinic, data from a real surgery environment is also collected – albeit in a limited amount. The process of data acquisition progresses in complexity, moving step by step towards realistic conditions, starting from artificial tissue models and advancing to human brain tissue. This progression is depicted schematically in Figure 4.1. Initially, data are gathered using artificial tissue models in a controlled environment, employing a computer numerical control (CNC) machine without tissue removal. This is followed by data collection involving tissue removal under CNC-controlled conditions. Subsequently, data are obtained in a handheld manner on artificial tissue models. Finally, data collection occurs in the operating room during neurosurgical procedures. The collected data serve as the basis for model training and evaluation, facilitating the development of classification approaches for distinguishing tissue stiffness and detecting contact states.

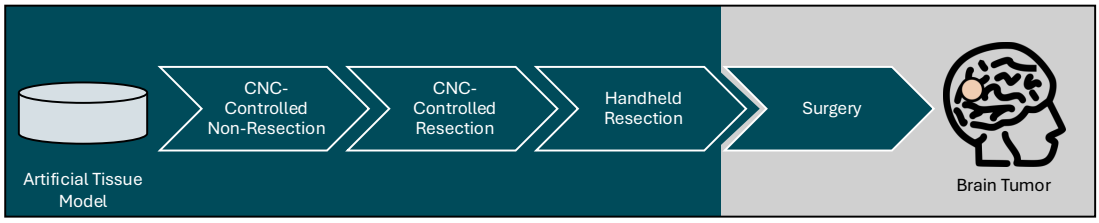


Figure 4.1: Flowchart depicting the progression of data acquisition. It commences with artificial tissue models in a controlled laboratory setting and advances to human brain tissue during surgery, with complexity intensifying at each stage. Gathered data serve for training and assessing models regarding tissue stiffness and contact state detection.

4.1 Tissue Types

4.1.1 Artificial Tissue Models

Since real biological tissue is difficult to obtain, artificial tissue models are manufactured to resemble the properties of brain tissue as closely as possible. This allows the data collecting process to be performed in a laboratory environment, where the tissue properties can be controlled, and the data acquisition process can be automated. Thus, the data collection process can be sped up by a larger factor which would not be possible in a real surgery environment where the data acquisition process can take months to years to obtain a sufficient amount of data.

Manufacturing Process

Manufacturing of the artificial tissue models is performed with a proprietary process by Söring GmbH in-house, thus cannot be disclosed in detail. The process is based on a polyvinyl alcohol (PVA) mixture and can be adjusted to match the desired properties. In general, the more PVA is added to the mixture, the stiffer the resulting tissue model becomes. Accordingly, the material becomes softer the less PVA it contains.

The PVA mixture is poured into multiple silicon molds and subsequently subjected to a freezing, thawing, and cooling process. The molds are cylindrical, with a diameter of 3 cm. The resulting tissue models are stored in bidistilled water to prevent dehydration and maintain their mechanical properties. In the end of the process, multiple tissue models are manufactured – each one can be referred to with its unique sample ID. All tissue models which are manufactured with the same PVA mixture are referred to with the same batch ID. Different tissue models differing in their stiffness are shown in Figure 4.2.

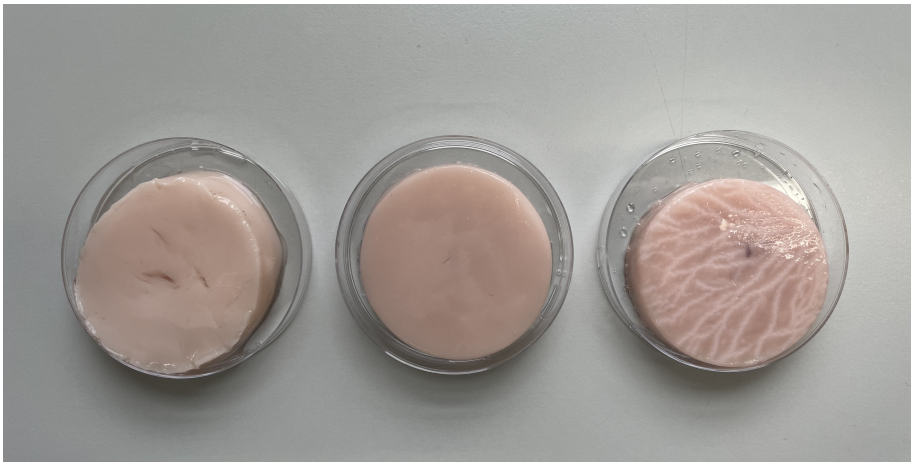


Figure 4.2: Different artificial tissue models. From left to right the tissue models are getting stiffer: 1000 Pa, 3000 Pa, 10 000 Pa.

Material Properties

The material properties of the artificial tissue models are measured in terms of stiffness which is captured with the mechanical property Young's modulus. This property measures the stiffness of a material and is defined as the ratio of stress to strain. Though, many different meachanical properties of a material can be measured, such as shear modulus, Poisson's ratio, etc., this work focuses on the Young's modulus as the main property due to feasibility of measurements in both ex vivo human brain tissue and artificial tissue models. Nevertheless, no implication is made that the Young's modulus is the only relevant property for the differentiation of tissue types. Answering this question would require a more comprehensive study which is beyond the scope of this work.

The Young's modulus is measured with an unconfined compression test. In this test, the tissue model is compressed in axial direction and can expand in radial direction. A tissue model is placed in a CNC machine which is equipped with a force sensor. Compression is performed with a cylindrical probe. Due to the similarity of the tissue models to real brain tissue, biphasic material properties are assumed. Thus, water is pressed out of the tissue model during compression, resulting in a non-linear behavior. To account for this, the tissue model is compressed in multiple steps with a waiting period of two minutes in between. This allows the tissue model to restore its original state with a stress-relaxation process [36].

For each batch ID, only a subset of the manufactured tissue models are measured. The average Young's modulus of the subset is used for all remaining tissue models of the same batch ID. Even though the manufacturing process is highly controlled, the tissue models still show a variance in their stiffness. This can be attributed to heterogeneities in the mixture and the freezing and cooling process, specifically local temperature gradients within the freezer and cooling unit. Thus, the assigned Young's modulus for each tissue model inherently contains uncertainty regarding the actual stiffness. In Figure 4.3, the assigned average Young's modulus and its corresponding standard deviation for each batch ID are shown. It can be seen that the variance increases with increasing stiffness. In order to remove batch IDs from the collected data that show a high standard deviation in dependence of the measured Young's modulus, a quadratic fit is performed on the data (red line in Figure 4.3). An upper bound based on the fit is used to remove batch IDs which are above this upper limit. The upper limit is calculated as

$$\text{upper limit} = 200 + 1.5 \cdot \text{quadratic fit} \tag{4.1}$$

which is empirically determined and shown as the dashed orange line in Figure 4.3.

4.1.2 Ex and In Vivo Tissue

Even though the main focus lies on artificial tissue models, data from real tissue is important for verification of the validity of the recorded data. Therefore, standard

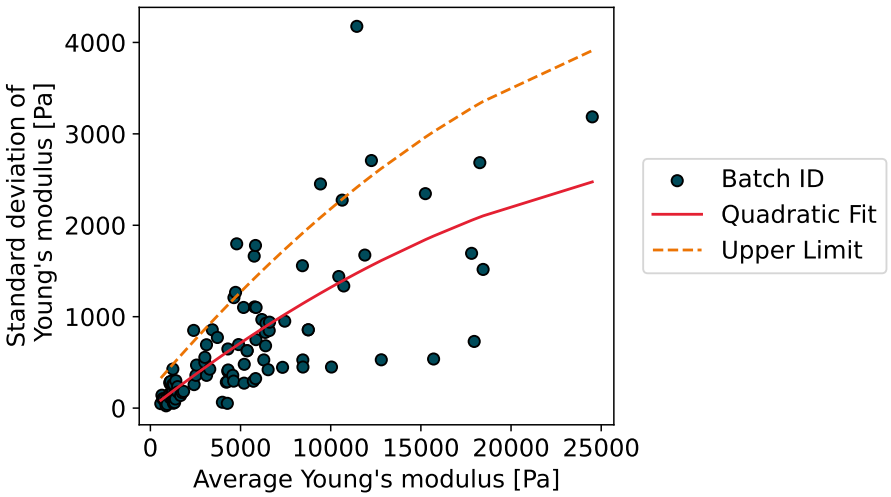


Figure 4.3: Average Young's modulus and standard deviation for each batch ID. The red line shows a quadratic fit to the data. The dashed orange line shows the upper limit for the standard deviation. Batch IDs above the upper limit are removed from the data.

operating procedures (SOPs) are developed from the clinical team to collect data from real tissue in a surgery environment when the ultrasonic aspirator is used.

Material Properties

Part of this SOP is the measurement of the stiffness of the tissue. However, due to the development of mechanical tests for different materials, necessitating specific test setups and loading conditions that exceed the limits of human brain tissue and may risk patient harm, stiffness is not measured in vivo but ex vivo [63, 101, 25]. If during the surgery biopsies are possible, tissue samples are taken with a minimum diameter of 5 mm and measured with an indentation device. For this Young's modulus measurement process, multiple measurements are performed on the same tissue sample with a loading of 0.3 g which is held for 30 s. Afterwards, tissue samples are diagnosed by a neuropathologist and the tissue type is determined. As a consequence, there is a discrepancy between the stiffness values and the data acquisition process, because the stiffness measurements are performed on ex vivo tissue samples, whereas the data acquisition process is performed on in vivo tissue. It cannot be guaranteed that the tissue type is the same for both processes. An overview of the average stiffness measurements and tissue types are shown in Figure 4.4. The tissue types are differentiated by the category of meningioma, glioma, and metastasis as encountered during surgery.

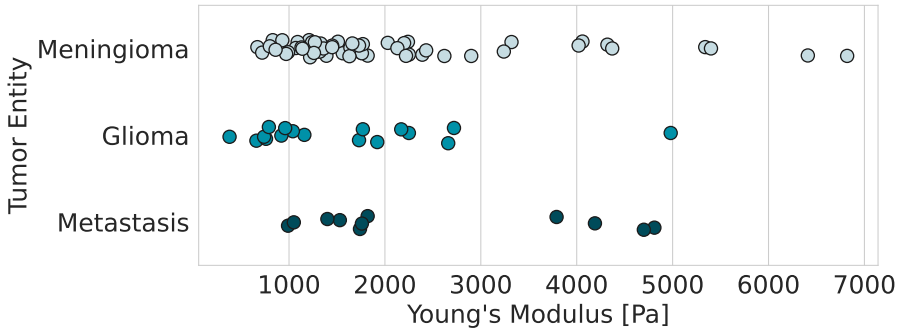


Figure 4.4: Young's modulus measurements and tissue types. Each dot represents a tissue sample measurement. Different colors indicate different tissue types.

4.2 Data Acquisition and Parameters

Data acquisition is performed with the ultrasonic aspirator. Specifically, the electrical data that is generated in the high frequency generator is recorded. In general, a differentiation between two high frequency generator types is made. One is the Gen_{prod} which has already been released to the market and is commercially available.[1] The other is the Gen_{dev} which is in development and not yet commercially available. For Gen_{prod}, data generation is performed externally between the generator and the ultrasonic aspirator handpiece using a measurement box. For Gen_{dev}, the data generation is performed within the generator itself, leveraging the internal data acquisition capabilities via controller area network bus. Throughout the process of this thesis, multiple changes to Gen_{dev} were made due to ongoing development of the device by Söring GmbH. Therefore, the recorded data parameters and conditions are not constant throughout the data collection process. Additionally, changes to the data acquisition process were made due to knowledge gained during the process. Necessary information about changes will be given in the according sections. Data sets will be referred to as

$$\text{D}i, \quad i \in \{1, \dots, 9\} \tag{4.2}$$

where i denotes the data set number which will be described in detail in the following sections. An overview of the data sets is given in Table 4.1.

4.2.1 Laboratory Environment

In a laboratory environment, the data acquisition process is performed either with a CNC machine or handheld on artificial tissue models. Both generator types are used in the process.

1 Sonoca 300, Söring GmbH, Quickborn, Germany

Table 4.1: Description of data sets used in this work.

Data	Description
D1	Non-resection data acquisition with Gen_{prod} with CNC
D2	Resection data acquisition with Gen_{dev} with CNC
D3	Contact classification with Gen_{dev} with CNC
D4	Comparison data acquisition with Gen_{prod} with CNC
D5	Manual data acquisition with Gen_{dev}
D6	Manual data acquisition with modified cover with Gen_{dev}
D7	Force influence data acquisition with Gen_{dev}
D8	In vivo data acquisition on human brain tissue with Gen_{prod}
D9	Human performance comparison study

CNC Machine

The same CNC machine used for the Young's modulus measurements is used for the data acquisition process. The instrument of the ultrasonic aspirator is mounted to the CNC machine and the tissue model is placed underneath the instrument. The angle of the instrument is set to 90° relative to the tissue model. A general acquisition with a CNC machine consists of three phases: 1) The instrument is lowered to the tissue model until it touches the tissue. 2) The instrument is moved in a lane profile over the tissue model and back while in contact with the tissue. 3) The instrument is lifted from the tissue model. During all three phases, the generator is active and the electrical data is recorded. During the contact phase, the instrument is constantly lowered down to a predefined penetration depth which is reached at the end of the phase. This way a constant contact is ensured. Due to the nature of CNC machines, information about the contact and non-contact state is available for each data point. Per tissue model, a total of five lanes are performed. It is ensured that the lanes are not performed on the same spot on the tissue model, and they have sufficient distance from the tissue model edges to prevent edge effects regarding mechanical properties.

Non-Resection[1] As an initial step for feasibility studies, the data acquisition process is performed on tissue models without the resection of tissue. This means that during the process, no tissue removal is performed. To this end, the generator is manipulated on a software level to the lowest possible ultrasound amplitude and disabled irrigation and aspiration. The generator used for this process is Gen_{prod}. Additionally, the sonotrode used for the data acquisition process is not equipped with an aspiration channel which deviates from the standard sonotrode used in neurosurgery. Data generated under

1 Part of master thesis of Denise Kesslau [92]

these conditions are referred to as D1. Captured electrical parameters are: voltage, current, frequency, active current, active power, power, and impedance. The sampling rate is 21 Hz. Four different tissue model types are considered during data acquisition which differ in their stiffness. Average Young's modulus of the tissue models range from soft to firm: 725 Pa, 2190 Pa, 5905 Pa and 20 829 Pa. With a recording length of 17.3 ± 2.7 s per lane, a total of 11 054 data points are recorded.

Resection After the initial feasibility study, the data acquisition is then further adapted to more realistic conditions: tissue removal is carried out during the acquisition with the CNC machine. Accordingly, usual generator settings are selected (ultrasound: 11 %, aspiration: 160 mbar, irrigation: 7 ml/min) and a standard sonotrode with an aspiration channel is utilized, which is commonly employed in neurosurgical procedures. Since the generator used for this data acquisition is Gen_{dev}, a higher sampling rate of 1000 Hz is possible. Nine electrical parameters are captured with this type of generator: voltage, current, frequency, active current, active power, impedance, phase, DC voltage, and resistance. Two tissue model types are utilized for this data acquisition process: 1513 Pa and 5709 Pa. At a recording length of $\sim$12.7 s per lane, this results in over 1 016 000 data points in total. The data recorded with this process are referred to as D2, given the focus of tissue differentiation in the context of resection.

Similarly, with the focus on establishing a data set for contact classification purposes, the same data acquisition process is performed with three different tissue models with average Young's modulus of 1513 Pa, 5709 Pa and 8460 Pa and four distinct generator settings, covering a variety of commonly used settings. Those generator settings are: 1) ultrasound: 11 %, aspiration: 160 mbar, and irrigation: 7 ml/min; 2) ultrasound: 29 %, aspiration: 280 mbar, and irrigation: 7 ml/min; 3) ultrasound: 50 %, aspiration: 360 mbar, and irrigation: 7 ml/min and 4) ultrasound: 79 %, aspiration: 620 mbar, and irrigation: 7 ml/min. In the context of the data set known as D3, these settings are denoted as *low*, *moderate*, *intermediate*, and *high*, respectively. A total of 2 935 295 data points are recorded with this process.

One data set is additionally recorded with the Gen_{prod} generator to enable a comparison with clinical data. Suitable settings of ultrasound: 32 %, aspiration: 150 mbar and irrigation: 4 ml/min are used on a variety of tissue models with Young's modulus values under 2000 Pa: 893 Pa, 1036 Pa, 1076 Pa, 1229 Pa, 1276 Pa, 1301 Pa and 1521 Pa. About 13 893 data points are recorded with this process. This data set is referred to as D4.

Handheld

In order to further adapt the realistic conditions of the data acquisition to the clinical conditions, the data acquisition is carried out in the laboratory in a handheld manner

while resecting tissue. This allows the capture of more data in less time, since the process is not limited by the CNC machine and the time it takes to perform the lane profile. Generally, the user brings the instrument in contact with the tissue model for a short period of time and then lifts it up again. This process is not limited to a lane profile or specified angles, but can rather be performed in any direction in a dabbing motion. In doing so, numerous unique contacts can be achieved on the same tissue model instead of being limited to five lane profiles. Throughout the data acquisition process, three different individuals perform the data acquisition. Data is recorded with Gen_{dev}, a 750 Hz sample rate and with the two different settings being 1) ultrasound: 31 %, aspiration: 180 mbar, and irrigation: 6 ml/min and 2) ultrasound: 78 %, aspiration: 210 mbar, and irrigation: 6 ml/min. Similar to the CNC machine data acquisition process, nine electrical parameters are recorded: voltage, current, frequency, active current, active power, impedance, phase, DC voltage, and resistance. Tissue models with an average Young's modulus of 1359 Pa, 3545 Pa and 5835 Pa are used. In total this results in 221 820 data points for all acquired data. As this data acquisition process is performed with the original cover component of the ultrasonic aspirator, the captured data are referred to as D5.

Due to the geometry of the original cover, there is a buildup of fluid that affects the signals. Thus, a new modified cover component was developed and applied. Data captured with this modified cover component are referred to as D6. The same data acquisition process regarding the sample rate and parameters is performed. Generator settings are: 1) ultrasound: 20 %, aspiration: 140 mbar, and irrigation: 4 ml/min; 2) ultrasound: 45 %, aspiration: 280 mbar, and irrigation: 7 ml/min and 3) ultrasound: 95 %, aspiration: 650 mbar, and irrigation: 14 ml/min. These settings are respectively referred to as *low*, *medium* and *high*. A variety of tissue models are used with Young's modulus ranging from 639 Pa to 15 705 Pa. An overview of the distribution of used tissue models per setting is shown in Figure 4.5 with its quantitative details in Table A.1. More than 82 447 000 data points are recorded with this process.

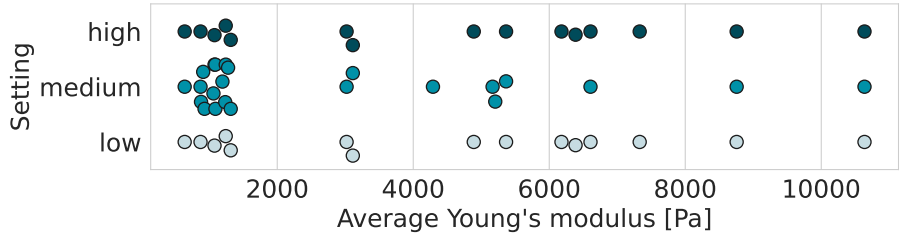


Figure 4.5: Overview of Young's modulus distribution per setting in tissue models used for the data acquisition process with a modified cover component D6. Each dot represents a batch ID. Different colors indicate different settings.

With this manual data acquisition process, information about the contact state of each data point is not available. For this reason, an offline contact detection algorithm

is applied to the data. This algorithm is based on the assumption that the voltage is higher during the contact state than during the non-contact state. With an analysis of the convolution of a step function and a peak detection on the voltage signal, the contact state can be determined.

In order to investigate the influence of force, data acquisition is accompanied by the use of a force sensor. The tissue model is put in a silicone mold which is placed on top of a force sensor as shown in Figure 4.6. Sponges are used for calibration between the electrical and force data by applying a relatively large force to the sponge. In a subsequent data processing step, the force data is synchronized with the electrical data by detecting the peaks in the force data and aligning them with the peaks in the electrical data. These peaks are used as synchronization points and are removed from the data. Force data is sampled with a 6.4 Hz sample rate which is resampled to match the electrical data sample rate of 1000 Hz. Tissue models with a Young's modulus of 915 Pa, 967 Pa, 18 000 Pa and 19 281 Pa are used on three different settings: 1) ultrasound: 19 %, aspiration: 140 mbar, and irrigation: 3 ml/min; 2) ultrasound: 46 %, aspiration: 290 mbar, and irrigation: 8 ml/min and 3) ultrasound: 100 %, aspiration: 740 mbar, and irrigation: 18 ml/min. The settings are again being referred to as *low*, *medium* and *high*. A total of 2 449 199 data points are recorded with this process on a total of 29 recorded sequences ranging in length from 36 s to 3.8 min. The user is asked to perform a similar force per sequence in terms of little or high force application. Data recorded with this process are referred to as D7.

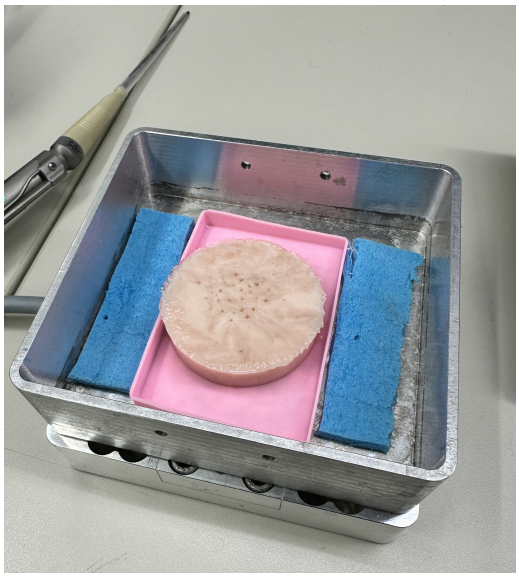


Figure 4.6: Force sensor setup for data acquisition process. Tissue model placed in silicone mold on top of the force sensor located inside a metalic housing. Blue sponges are used for synchronization between electrical and force data.

4.2.2 Operating Room Environment

In the operating room, data acquisition is performed by a neurosurgeon during a neurosurgical procedure in vivo. The generator used for this process is Gen_{prod} with a sample rate of 21 Hz and settings depending on the procedure and the subjective preference of the neurosurgeon. The same electrical parameters are recorded as in the non-resection data acquisition process with the CNC machine (paragraph 4.2.1): voltage, current, frequency, active current, active power, power, and impedance. The SOP for data acquisition in the operating room is as follows: When the ultrasonic aspirator is indicated, the neurosurgeon is asked to perform ten individual contacts to the tumorous tissue with the instrument. If possible, biopsies are taken from the tissue. Afterwards, the usual operation continues to avoid unnecessary interference with the surgery. As mentioned in subsection 4.1.2, the stiffness of the tissue is measured ex vivo with an indentation device and cannot be guaranteed to be the same tissue type as the in vivo tissue. Thus, a direct connection between the stiffness measurements and the recorded data is difficult. A total of 26 surgeries are performed with this SOP resulting in 10 868 data points. Data recorded in the operating room are referred to as D8.

4.3 Data Labeling

With the stiffness being the target variable for differentiating between tissue types, the data needs to be labeled accordingly. Due to the nature of the data acquisition process, the stiffness indicated by the Young's modulus is given as a mean value for all samples of the same batch ID. Additionally, the variance of the Young's modulus is provided. Classifying the data into tissue types solely based on the mean value of the Young's modulus is not feasible, since the variance of stiffness is not considered. This can lead to misclassifications, since the stiffness of a tissue model can vary within the same batch ID.

Therefore, binning of the data is performed to account for the variance of the stiffness. To this end, an optimal binning algorithm is used to determine the optimal bin edges for the data [121]. The continuous numerical target for the binning is the standard deviation of the Young's modulus, with the mean value of the Young's modulus as the variable to be discretized. In a first step, a pre-binning is performed with a regression decision tree which allows for a rough estimation of m initial split points. The resulting splits are sorted by ascending order and create $n = m + 1$ initial bins. Subsequently, refinement of the initial bins is performed with a mixed integer programming formulation under constraints. Constraints being that initial bins stand alone or form a broader bin interval, but cannot be eliminated, only initial bins that are consecutive can be merged, with the last interval reaching ∞. Additionally, limitation in maximum number of bins, minimum number of samples per bin, monotonicity and maximum p-value constraints between bins can be employed. Assessing the quality of the binning can be performed with the

Information Value (IV) as

$$\text{IV} = \sum_{i=1}^{n} |\mu - u_i| \frac{r_i}{r_T}, \tag{4.3}$$

where μ is the global mean of the target variable, u_i and r_i are the mean target variable and number of samples of the i-th bin and r_T is the total number of samples. The interested reader is referred to [121] for extensive details on the mathematics of the optimal binning algorithm.

On the basis of the optimal binning algorithm, the data is binned into a maximum of seven bins with a minimum of 15 % of the total number of samples per bin. The resulting bins are shown in Figure 4.7 with their bin edges and number of samples per bin. Bin edges in Pa are [0, 1090), [1090, 2473), [2473, 4754), [4754, 8755), [8755, 25 000], while $-\infty$ and ∞ are set to the available minimum and maximum values of the Young's modulus. Bins are labeled with the midpoint of each bin, rounded to the next hundred, as shown in Table 4.2. Within the scope of this work, those bins are referred to as Bin_{opt}. It is important to emphasize that the data utilized for binning is not solely confined to the information collected for electrical data acquisition. Rather, it encompasses a broader range, including additional Young's modulus measurements that are available.

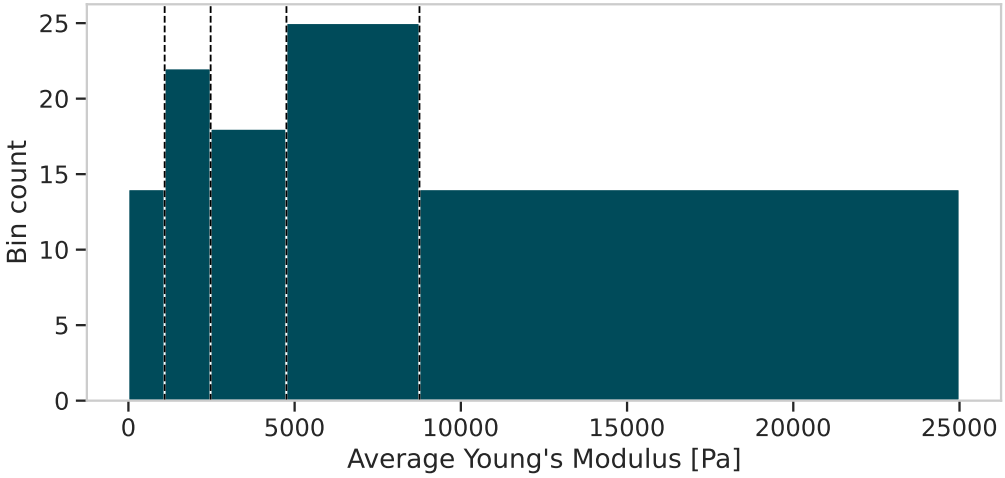


Figure 4.7: Optimal binning of the data with bin edges shown as dashed vertical lines as per Bin_{opt}. Bin count indicates number of batch IDs per bin.

In the context of hierarchical classification, data binning is applied in a hierarchical, binary manner with a minimum of 30 % of the total number of samples per bin. On a first level the data is discretized into two bins, named as *soft* and *firm*, with the bin edges being [0, 5844), [5844, 25 000] in Pa. This binning is referred to as Bin^1_{hier}. On a second level, the *soft* bin is further discretized into two bins, named as *very soft* and *medium soft*,

Table 4.2: Bin edges and labels of Bin_{opt}

Bin Edges	Label
[0, 1090)	500
[1090, 2473)	1700
[2473, 4754)	3600
[4754, 8755)	6700
[8755, 25 000)	16 900

with the bin edges being [0, 2473), [2473, 5844) in Pa. Due to the lack of sufficient data, the *firm* bin is not further discretized. Accordingly, this binning is referred to as Bin^2_{hier}. An overview for the bins and the associated labels per level is found in Table 4.3.

Table 4.3: Bin edges and labels per level for Bin^1_{hier} and Bin^2_{hier}. Level 2 labels for last bin edges not applicable.

Bin Edges	Level 1	Level 2
[0, 2473)	*soft*	*very soft*
[2473, 5844)		*medium soft*
[5844, 25 000]	*firm*	-

4.4 Human Tissue Differentiation Capability

In order to quantitatively compare the performance of the classification models to the human's capability, a human performance comparison study is conducted.[1] To this end, 36 participants are asked to classify tissue models in the classes *soft*, *medium*, and *firm* in a blinded manner. Of the 36 participants, 13 are neurosurgeons which are considered as experts, the remaining 23 are considered non-experts due to their lack of medical background. The 13 experts are further divided into two groups: seven are considered as highly experienced (specialists) and six as less experienced (assistant physicians). Classes are renamed for easier understanding and represent the previously used classes of the hierarchical labeling of Bin^1_{hier} and Bin^2_{hier}, *very soft*, *medium soft*, and *firm*, respectively. Before the blinded classification of tissue models, the participants are given a training phase in which a representative tissue model per class is shown. The participants are allowed to touch the tissue with their hands with latex gloves on and/or use forceps to get a feeling for the classes. The tissue remains visible to the participants during the training phase. This training phase has no time limit and is only

1 The ethics committee did not consider any consultation on the study to be necessary. No ethical concerns were raised.

finished when the participants feel confident in their classification capability. In a way this setup is similar to the training and test phase of a classification model. Following the training phase, participants are instructed to classify nine tissue models sequentially in a blinded manner during the test phase. The total quantity and class distribution of tissue models are undisclosed to participants, and they are presented in a randomized sequence. No feedback is provided to participants throughout the test phase. Moreover, tissue models utilized for training purposes cannot be referenced during the test phase, nor can previously classified tissue models. The setup is depicted in Figure 4.8. On the bottom left the testing area is shown, where the tissue models are presented to the participants in a blinded manner by using a black curtain. The training area is shown on the top right, where the tissue models are presented to the participants in an unblinded manner where they can touch the tissue models and familiarize themselves with the classes.

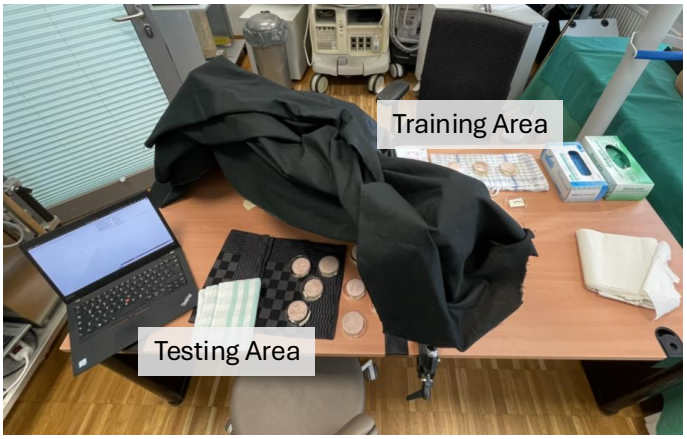


Figure 4.8: Setup of the human performance comparison study. The training area is depicted in the upper right corner, where participants are exposed to tissue models in an unblinded fashion. The testing area is illustrated in the lower left corner, where participants interact with tissue models in a blinded manner behind the black curtain.

In total, twelve tissue models are used throughout the study, three for the training phase and nine for the test phase. An overview of which batch IDs are used for the training and test phase, how many tissue models per batch ID, which class they are assigned to and the associated Young's modulus is shown in Table 4.4. As depicted in the table, batch IDs F-C, M-B, and S-A are utilized both during training and testing phases. However, the tissue models employed during training are distinct from those utilized in the testing phase. This choice arises from the constraints of the study involving human participants and practical limitations. Consequently, only a restricted number of tissue models can be utilized. Therefore, the decision to employ tissue models with the same batch ID for both training and testing phases is deliberate, aiming to ensure

that participants familiarize themselves with class distinctions, while the selected tissue models remain representative of the respective classes during testing. It is important to note that while tissue models from the same batch ID are selected, they may not possess identical tissue properties due to variances inherent in the production procedures. The data collected from the human performance are referred to as D9.

Table 4.4: Overview of batch IDs and their associated Young's modulus, classes, amount and study phase for the human performance study.

Batch ID	Class	Model Amount	Study Phase	Young's Modulus [Pa]	
				Mean	Standard Deviation
S-A	Soft	2	1× Train/1× Test	911	105
S-B	Soft	2	Test	1633	482
M-A	Medium	1	Test	4266	644
M-B	Medium	3	1× Train/2× Test	2742	730
F-A	Firm	1	Test	16 877	609
F-B	Firm	1	Test	7842	600
F-C	Firm	2	1× Train/1× Test	10 686	1066

4.5 Discussion

Acquisition of electrical data with an ultrasonic aspirator is performed on different tissue types, environments and generator types. The data acquisition process is mostly focused and accompanied by constant further development on the generator Gen_{dev} without any influence and without exact knowledge of internal changes. A systematic check for impact due to changes is not possible due to the time and resource intensive nature of the data acquisition process. Moreover, tissue is destroyed during the resection process, which renders it impossible to replicate the data acquisition process. Thus, the entire data needs to be considered as ambiguous with unknowns regarding the exact conditions of the data acquisition process.

With three different settings of the generator, which can be selected independently, a large number of different combinations is possible. However, not all combinations are useful for the resection process. The settings should ensure an efficient resection of the tissue. The used combinations throughout this thesis are evaluated to have an efficient resection result. However, the subjective feeling of optimal settings for proper resection does not necessarily correlate with the actual resection results. Internal studies by Söring GmbH observed that surgeons often set higher settings for the resection process, which does not appear to result in faster or better outcomes. This difference in choice of settings can be observed between the data sets D8 and D6 in later sections of this thesis.

One additional factor that influences the choice of settings is the type of tissue. Not all tissue types can be resected with the same settings. For example, resection on firm tissue types with low settings barely results in any resection, whereas resection on soft tissue types with high settings results in too much resection which can lead to uncontrollable fragmentation of the tissue. Thus, trade-offs between settings and tissue types are made throughout the data acquisition process in the laboratory environment. This means that some acquisitions are performed with settings that do not necessarily result in an efficient resection process, but need to be considered for a comprehensive data set.

Another factor to consider is the force applied by the user during the data acquisition process. During CNC machine data acquisition, the force applied by the machine is set by the process to not exceed more than 40 mN. This value is based on internal studies by Söring GmbH with neurosurgeons showing an average force of 30 mN on tissue model experiments which match the clinical conditions as closely as possible. Due to the very soft nature of brain tissue and the risk associated with the surgery, neurosurgeons are trained to apply a low force. Similarly, the force applied by the user during handheld data acquisition in the laboratory environment is kept as low as possible, without explicitly measuring the force. A more detailed analysis of the force applied by the user during the data acquisition process is provided in later sections of this thesis.

Direct comparability between the electrical data of the laboratory environment and the operating room is limited due to the different generator types (Gen_{prod} and Gen_{dev}) and the different environments. Using Gen_{dev} in the operating room is not possible due to the ongoing development of the generator and the lack of approval for clinical use. Nevertheless, captured data in the operating room allowed to gain insights the workflow of the data acquisition process during surgery. Due to the limitations of Gen_{prod} and the behaviour of the neurosurgeons, data was only captured while the instrument was already in contact with the tissue with a low sampling rate of 21 Hz. To account for that, the SOP was adapted to explicitly measure ten contacts to the tissue with the instrument, which allowed to record data usable for analysis.

Furthermore, the complexity and limitations of the tissue manufacturing process must also be considered in the overall context. Due to its inherent fluctuations, the manufacturing process is only capable of producing exact, desired tissue stiffness with a certain level of accuracy. Thus, closing gaps in the stiffness range remains difficult. These variations in stiffness within a batch also mean that it is only possible to assign tissue classes to the data to a certain extent and exactness. A more detailed and precise categorization and determination of tissue stiffness is not currently available or feasible. However, other mechanical properties of the tissue models, such as the tissue relaxation, could be considered in future work to further differentiate between tissue types.

Data collected from the human performance study (D9) are utilized to compare the performance of classification models with human capability. This comparative analysis is crucial for assessing the potential applicability of the models in clinical settings and

for establishing a performance baseline they should meet. Insights from the human performance study offer valuable information for the advancement of classification models and for guiding future research. Coupled with laboratory data that closely mimics clinical conditions, these initial findings help establish a correlation between tissue and electrical data, paving the way for the development of an intelligent ultrasonic aspirator.

CHAPTER 5

Contact Detection

For a human operator, assessing if an ultrasonic aspirator is in contact with tissue when performing tissue resection is a natural task, since a human can simply see, feel and hear the contact. However, in the context of an automated process, this information is not trivial to obtain but serves as a crucial input for further processing steps. Within the scope of differentiating tissue types, the information of contact is essential due to two reasons. Firstly, the process of classifying tissue stiffness is built in a hierarchical manner as seen in Figure 5.1. Meaning, the classification of tissue types is only permitted once contact is established. This approach allows for the process to be highly controlled and manageable, as it focuses on a single task at a time. Moreover, this approach allows for the reduction of the necessity for complex models to differentiate between tissue types and additional cases such as *no contact*. Secondly, feature generation processes benefit from the knowledge of the given information regarding contact establishment, allowing for a more focused feature generation process. Additionally, other processes outside of tissue classification can benefit from the knowledge of contact establishment in different subsequent tasks, which are discussed at the end of this chapter. Thus, this chapter is investigating the feasibility of detecting contact between the ultrasonic aspirator and tissue to answer research question I) introduced in chapter 1. The placement of this task in the overall context of the intelligent ultrasonic aspirator is shown in Figure 5.1 and highlighted in blue. The chapter is initially investigating the contact detection in a controlled environment utilizing a CNC machine while tissue resection is performed. In addition to contact detection, the real-time capabilities of the methods are investigated to ensure the usability of the methods in a real-world application. In a subsequent step, the complexity of the task is increased by introducing a more realistic environment in which the ultrasonic aspirator is controlled in a handheld manner, mirroring the conditions of real surgery.

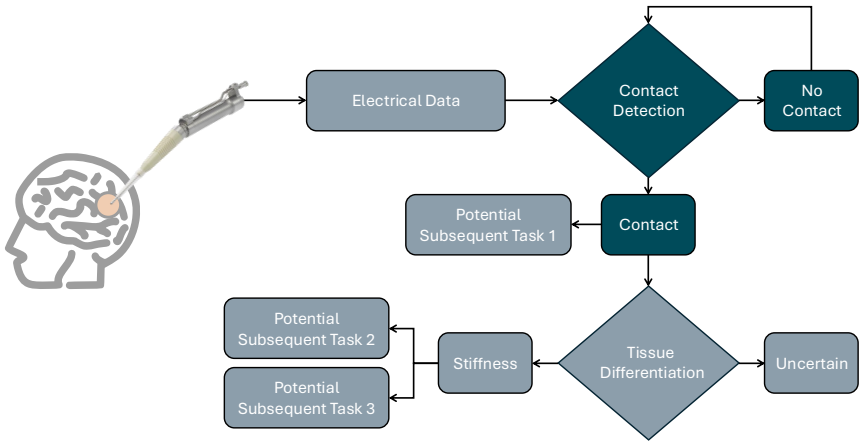


Figure 5.1: Flow chart of an intelligent ultrasonic aspirator. The content of this chapter is highlighted in blue.

5.1 Related Work

Identifying the contact of a surgical tool under tissue removal conditions and soft tissue presents a novel challenge, without any direct comparison available to the best of the author's knowledge. However, within the broader scope of identifying interactions between tools and objects across various domains, numerous methodologies have been proposed. Specifically, in the simulation domain, diverse approaches have been suggested for detecting contact interactions between a tool or user and tissue. For instance, in the domain of virtual reality surgical training, accurate contact detection is imperative for providing realistic haptic feedback and accurately representing tissue deformations [102, 83]. These deformations can be replicated through methods such as employing a mass-spring model [102], utilizing position-based dynamics [70], or employing finite element models [39, 86], which can accommodate multiple topological changes arising from cutting, contact, or other interactions in real-time. As these approaches primarily rely on simulations, contact detection can be simplified by focusing on collisions between the geometries or surfaces of the objects, rather than relying on specific sensor data.

While not directly related to interactions between tissue and surgical instruments, analogous approaches employing similar foundational data structures can be found in other disciplines. For instance, in the gait analysis work of Kidziński et al. [93], event detection is utilized to automatically identify foot-contact and foot-off events in children with both normal and pathological gait patterns. In this study, the authors employed a recurrent neural network architecture to analyze kinematics and marker trajectories, followed by a peak detection algorithm to identify events with average errors ranging

from 10 to 13 ms.

In the realm of tissue and instrument interactions, Yoshimoto et al. [183] introduced a smart sensing system capable of estimating tissue behavior and user actions. This system incorporates a sensor utilizing electrical contact impedance via resistive coupling, specifically designed for integration with surgical tools. When the sensor makes contact with the tissue, the circuit closes, enabling the measurement of voltage and subsequent data processing. Consequently, contact detection can be achieved through straightforward thresholding, facilitating the differentiation between initial contact and manipulation phases. This enables the identification of retracting, pushing, and pressing actions to a certain extent, proving particularly valuable in minimally invasive surgeries where direct tissue feedback is lacking.

Motivated by this, Sühn et al. [168] developed an acoustic emission sensing tool tailored for arthroscopic surgery. Their tool, comprising a conventional spring-loaded Veress needle coupled with a microphone, captures and analyzes the propagated acoustic signals during palpation at a sampling rate of 44.1 kHz. Contact detection with explanted human femoral heads was achieved by computing the envelope of the signal via continuous wavelet transformation using a Morlet wavelet, focusing on scales corresponding to pseudo-frequencies around 4 kHz. Subsequent spectral analysis of a windowed signal, surrounding the detected contact events, provided preliminary qualitative insights into the feasibility of this approach.

Acoustic emission sensing is also prevalent in various machining techniques, where precision is paramount. Min et al. [117] evaluated contact detection based on acoustic emission signals while minimizing damage to the workpiece. Utilizing simple thresholding of the acoustic emission signal, they controlled the feed rate of the tool through two distinct approach methods. While complete elimination of damage remains elusive, Turner et al. [175] explored the use of acoustic emission sensors to measure the standoff distance between ultrasonic tools and workpieces in a non-contact manner to mitigate damage. Their study demonstrated that a fast Fourier transformation of the acoustic emission signal facilitated distance detection by analyzing the fundamental frequency and its harmonics. Notably, under contact, similar intensities of the harmonics and fundamental frequency were observed, while standoff conditions of less than 10 µm exhibited significantly lower harmonic intensities, enabling clear differentiation between contact and near-contact states.

Despite parallels to ultrasonic aspirators, the application of acoustic emission sensors in machining involves materials with distinct properties and operating conditions, complicating direct method transfer.

In the domain of robotic interactions, contact detection holds paramount importance in ensuring both the safety of the robot and its environment, as well as enabling various manipulation tasks. The transition between different manipulation phases, such as grasping, pushing, recognizing slippage and sliding, or exploring and identifying objects, often relies on effective contact detection mechanisms [89, 150, 31, 32].

Various sensors can be employed for contact detection, including force-torque sensors, joint angle sensors, or tactile sensors [77]. Sandykbayeva et al. [151] utilized tactile vibration sensors in their work, employing active vibration sensing through an eccentric rotating mass motor and accelerometer to detect contact between robotic fingertips and objects. They identified a specific frequency range in the accelerometer signal's spectrogram to detect contact events, with changes in peak amplitude indicating contact, utilizing a threshold set at twice the noise level. This approach demonstrated superior performance compared to force sensors, detecting nine out of a total of ten contact events.

Similarly, Dai et al. [43] investigated vibration-based depth estimation during bone drilling, utilizing accelerometers for vibration recording. They identified initial bone contact by detecting changes in harmonic amplitudes in the frequency domain.

In the realm of safety-critical evaluations involving robots and humans, collision detection is of utmost importance. Haddadin et al. [69] conducted analyses on soft tissue injuries caused by sharp tools mounted on robots in collision scenarios with silicone, porcine samples, and human subjects. They implemented a sensorless system for contact detection, employing a disturbance observer to estimate external forces on the robot's joints. When these forces exceeded predefined thresholds, indicating collisions, the robot halted immediately. In scenarios involving kitchen knives and scissors, this approach effectively prevented tool penetration. Without contact detection, the penetration could reach 100 %, whereas with the implementation of contact detection, penetration was effectively minimized, reaching zero.

Furthermore, in the context of assisted robotic surgery, contact detection, facilitating force feedback, proves crucial for surgeons to mitigate the risk of tissue damage [44, 171, 135]. To enhance automation in this domain, Psota et al. [134] introduced a visual contact detection system to identify contact between two robotic arms and tissue. The system employs a convolutional neural network based on the U-Net architecture [145] to classify contact and no contact on a pixel level. Input data consist of three grayscale images representing different time instances: 0.5 s before, 0.25 s before, and the present time. Results demonstrate high accuracy, up to 90 %, in detecting contact on various tissues such as beef and chicken.

For handheld tools, maintaining consistent contacts in terms of force is crucial for accurate measurements and applications. Latt et al. [104] developed an instrument for probe-based confocal laser endomicroscopy (pCLE) with steady tissue contact. The system, comprising an actuator, a force sensor, and a pCLE probe, maintains defined contact force using a closed-loop control system. Evaluation on artificial liver and porcine colonic tissue in bench and handheld settings showcased the system's ability to sustain constant contacts under disturbances, with root mean square errors of less than 6 mN and improved pCLE image quality.

Automated contact detection in handheld devices is also explored by Ruderman et al. [148], who utilized a fiber optic probe with polarization gating spectroscopy for

contact detection. Specific wavelengths exhibiting high reflected signal intensity were employed in a thresholding manner to detect contact. To ensure steady contact, five consecutive measurements allowed a maximum variation of less than 3 % in signal intensity; otherwise, no contact was assumed. In in vivo colonoscopy studies, the method achieved up to 91 % accuracy in contact detection and reduced user variability.

Direct transfer of these methods to ultrasonic aspirators presents challenges for several reasons. Firstly, the methods are tailored specifically to their respective applications in contact detection, which may not align with the requirements of ultrasonic aspirators. Secondly, the tissue materials used in these studies differ from those found in the human brain and may not encounter the same conditions present in softer domains [25]. Lastly, contact detection for ultrasonic aspirators faces the additional challenge of simultaneous tissue removal, which is not as prominent in other applications. Thus, making the automatic contact detection task with a working ultrasonic aspirator for very soft tissues a unique and never before addressed challenge.

5.2 Feasibility under Controlled Conditions with Tissue Ablation[1]

Data Preparation The data set D3 is utilized for contact classification purposes. Data are recorded under controlled conditions in a laboratory setting using a CNC machine while resecting tissue. Different generator settings for tissue removal are used throughout this data collection process: *low*, *moderate*, *intermediate*, and *high*. Due to the nature of the data collection process, data acquisition is done in a lane fashion on three different tissue types which are referred to within this study as *soft*, *medium*, and *firm* in terms of stiffness. In total, 172 recordings are available after a thorough data cleaning process, giving a total of about 2 184 400 data points to be classified. Since the data are recorded in a controlled CNC environment, precise ground truth labels are available when the ultrasonic aspirator is in contact with the tissue. This information is used to generate the ground truth labels for the contact classification task as *contact* and *no contact*.

Methods Four different data-driven methods are utilized to classify each time-series data point into either *contact* or *no contact*. Even though the data are pre-recorded, it is handled as online data. In other words, in each time step, a new data point arrives that has to be classified based on the current and previous values without prior knowledge of future values.

The first method is a change point detection algorithm (CPD). Since the premises of the data are that states change over time, this statistical approach is used as a baseline method. A change point detection is not a classification method in the sense of

1 Parts of this section are based on the publication [10] © 2023 IEEE.

distinguishing between *contact* and *no contact*, but a method to detect the change of states in a temporal context. To deduce which class the current new data point is to be assigned to, it is checked whether the state has changed. If this is the case, the opposite class since the last change point is used; otherwise, the same class remains. This is only possible because it is known that the starting point is always in the *no contact* class. Specifically, a Bayesian online CPD approach is employed using a constant hazard function $H = 1/\lambda$ as a change point probability prior with $\lambda = 6000$ [3]. Prior information about the data are modeled with a Gaussian model $\mathcal{N}(\tilde{x}|\mu_0,\sigma_0^2 + \sigma^2)$. The empirically determined data variance σ^2 is set to 300, while the prior mean μ_0 and the prior variance σ_0^2 are initialized with the mean and the variance of the first 100 ms of the current time-series, respectively. These first 100 ms can be seen as a warm-up phase at the beginning of each use in which no predictions can occur. Because the uncertainty of whether a change point is detected remains high with each new data point, (another) 100 ms are waited for with each change point prediction. This consequently results in a constant 100 ms delay in the online prediction, meaning the last 100 values cannot be detected as change points. The method relies only on the electrical parameter voltage.

The remaining approaches do not detect change points but classify a new data point directly as *contact* and *no contact*. All electrical parameters are utilized for online prediction purposes. A random forest (RF) is used, which provides a well-known and straightforward structure [17]. Since this is a data point-based method, i.e. classifications are only carried out based on the current data, no temporal dependencies exist in the prediction. The Gini index is used as feature importance with a tree size of 100.

To leverage the time dependence of the signals, the next method is a recurrent neural network (RNN) in the form of a long short-term memory (LSTM) model with 64 nodes [75]. This deep learning-based model uses data from a sliding window of size 75 for prediction purposes, which represents a time span of 75 ms. For better training performance and less overfitting, batch normalization and one dropout layer with a dropout rate of 0.3 are utilized. The final binary prediction is made by a dense layer of size one using a ReLU activation function.

The fourth and last time-dependent method utilizes a deep learning-based temporal convolutional network (TCN) structure [7]. By using causal convolutions and dilations, TCN can adapt to sequential data with temporal aspects and large receptive fields. For classification, a TCN layer with 64 filters, kernel size of three and dilations list of 2^i with $i = 0, \dots, 5$ steps is used. Similarly to the LSTM, it is based on a sliding window of size 75 and uses batch normalization, one dropout layer with a dropout rate of 0.3 and a final dense layer of size one using a ReLU activation function.

Because of the need for an online application, predictions for LSTM and TCN are always done for the last value in the sliding window which is shifted one data point at a time. This means the methods get the current information and 74 additional data points from the past for each prediction. The window size is determined empirically, such that it is as small as possible but preserves all information necessary for accurate contact

classification. Consequently, the first predicted value for LSTM and TCN is delivered after 75 ms. Both deep learning methods are trained with Adam optimization with a learning rate of 0.01. Binary cross entropy is used for loss calculation with a batch size of 256 and all neural networks are trained over 100 epochs.

As a first experiment, the general performance of the four contact classification methods CPD, RF, LSTM and TCN are evaluated. Here, the raw electrical data signals are processed without any preprocessing before classifying the time-series data on *contact* and *no contact*.

In a second experiment, the impact of two different data scaling preprocessing steps is analyzed. Since the data are used in an online application, it should not be necessary to know a priori in which numerical range the data will be during tissue ablation. However, since the measurement system, i.e. the ultrasonic aspirator generally behaves in a repeatable and predictable way, the values during the *no contact* phase can be estimated to some degree. This prior information is therefore used to scale the data so that *no contact* always takes the base level around zero.

Note that the scaling preprocessing approach is not performed for CPD since this requires only an adjustment of the prior probabilities, as the data are only shifted. Consequently, this does not affect the detection of change points.

The first scaling approach calculates its scaling values for each recording separately and is referred to as *live scaling*. For each electrical feature of a single recording, the mean values over the first 100 data points are calculated. Those values are the scaling values that are subtracted from each data point of the rest of the signal. To reduce overhead, those values are calculated prior to training. Due to the short time interval of 100 data points, the ultrasonic aspirator is not in contact with tissue yet. As a result, the calculated values correspond to the numeric range of the *no contact* area and scale it towards the zero level. Thus, with each recording, the system recalibrates itself.

The next scaling approach is based on a lookup-table that represents all known settings and is referred to as *lookup scaling*. It is assumed that the *no contact* phase is similar for each setting because, at this time, the instrument should constantly be exposed to the same condition – being in the air. For this, the mean of the initial *no contact* area (i.e. before contact is established) is calculated over all recordings for each setting. Similar to the first approach, those values are subtracted from the corresponding signals.

A third experiment examines the influence of smoothing the data as a preprocessing step. To ease the learning component of the machine learning algorithms and the data distribution for the statistical CPD approach, the high-frequency parts of all electrical signals are smoothed. A Butterworth low-pass filter is considered, which is a linear analog electronic filter that passes signals lower than an empirically determined frequency of 15 Hz. For filtering purposes, a sliding window is used – even for CPD and RF, where a sliding window does not contribute to classification. This means each window is filtered separately. For CPD and RF methods, the sliding window is adjusted so that only the

most recent data points within each window are considered for classification.

In the fourth subsequent experiment, the two preprocessing steps of scaling and smoothing filters are combined and the influence on the classification is investigated – again with the exception of CPD for previously stated reasons.

All experiments are conducted with five-fold cross-validation. Care is taken that all settings and tissue classes are represented in similar distribution in all data splits while data from the same recording are not divided. Since CPD is a statistical method with no explicit learning process, predictions are made only on the respective test data of the folds to remain consistent in the evaluations.

To prevent overfitting on deep learning-based methods, another split is performed on the training set to create a validation set. The distribution of the test, validation and training sets is approximately 60:20:20, respectively.

For results quantification, multiple classification metrics are calculated: accuracy (ACC), F1-score (F1), positive predicted value (PPV) and true positive rate (TPR). These metrics represent the overall performance of each model.

In addition, three explicitly developed metrics for this classification problem are used to better understand classification performance. First, the number of counted contact areas (CCA). Here, the ideal result is one because each recording has precisely one *contact* phase. Next, the time to the following contact area is used for the beginning (TT-S) and ending (TT-E) in milliseconds. This means the distance in time between the ground truth *contact* area and the nearest predicted *contact* area is calculated. The best achievable result is zero, i.e., the predicted areas have the same start point and endpoint as the ground truth *contact* area.

For clarification, all metrics are calculated as mean with standard deviation over all five cross-validation folds.

General Classification Performance The general classification results of the four presented approaches without data preprocessing are shown in Table 5.1. Each of the four methods demonstrates different performances. CPD's overall result ranks the highest with an F1 of 0.862 ± 0.052, closely followed by LSTM with an F1 of 0.838 ± 0.040. The RF method achieves 0.772 ± 0.018 in the F1 metric even though no temporal information is included in the prediction process. In contrast, the performance of TCN in F1 of about 0.555 ± 0.278 indicates about half of the data points are not classified correctly. In addition, the high standard deviation illustrates a greater dispersion over the cross-validation runs. Although in CPD, the CCA gives a very favorable result compared to the rest of the methods with 1.13 ± 0.04, at the same time, a comparatively high deviation from the actual start and end points of the *contact* in TT-S and TT-E of about 500 ms can be observed. On the other hand, the RF method shows a close proximity to the optimal start and end values TT-S and TT-E with 37.0 ± 21.8 ms and 30.9 ± 32.9 ms, respectively. However, this is apparently at the expense of a high average distinct *contact* number

CCA of more than 455.

A more detailed look at the metrics provides insight and understanding of performance. Of particular interest to note is that the TCN method performs well on CCA, while it generally performs poorly. This is mostly caused by this method's shortcoming of predicting all data points of an entire recording as the same class – either as *contact* or *no contact*. This is reinforced by the fact that the CCA value of this measurement should be zero if *no contact* is predicted for all data points and one if *contact* is predicted for all data points. The reason for the low performance is caused by the lack of data scaling, which as a consequence, hinders a sufficient learning process of the TCN.

Another discrepancy in metrics can be seen with CPD. Here, a good performance in CCA can be seen with a relatively large TT-S and TT-E deviation from the ideal value. In this case, this is due to the inherent behavior of the CPD method. This method can only detect change points. In a few cases, only the initial or only the final change point (beginning or end of the *contact*) is found. So as soon as a change point is not detected correctly, all following change points are not classified correctly. This is due to the fact that the classification is based on the change point detection and a detection shortly after a missed change point is unlikely. This is the biggest advantage of the other methods because a direct classification is possible at any time and works independently of the change points.

Performance Effect – Scaling As one of the preprocessing approaches, data scaling is used to reduce numerical variations between different recordings and electrical features to provide a similar baseline under different conditions. The results of the two scaling approaches are displayed in Table 5.2.

In general, improvement can be seen by employing either of the two scaling approaches. The highest results were obtained with TCN and *live scaling*, with an improvement of 0.422 in the F1 metric to 0.977 ± 0.007. This is most likely caused by the numerical change of the input data, which allows the network to learn sufficient features in the data. *Live scaling* performed reasonably well for both deep learning-based methods and RF. In a direct comparison of the two approaches, better performance can be seen for the deep learning-based methods in the F1 metric, for example. Here, the results

Table 5.1: Results of different contact classification methods. Metrics are stated as the mean and standard deviation over five cross-validation. The best value of each metric is highlighted.

Method	F1	ACC	PPV	TPR	CCA	TT-S	TT-E
CPD	**0.862** $\pm$ 0.052	**0.858** $\pm$ 0.052	**0.874** $\pm$ 0.051	**0.850** $\pm$ 0.054	**1.13** $\pm$ 0.04	494.0 $\pm$ 153.2	508.0 $\pm$ 206.7
RF	0.772 $\pm$ 0.018	0.759 $\pm$ 0.021	0.753 $\pm$ 0.026	0.772 $\pm$ 0.018	455.24 $\pm$ 24.37	**37.0** $\pm$ 21.8	**30.9** $\pm$ 32.9
LSTM	0.838 $\pm$ 0.040	0.824 $\pm$ 0.055	0.830 $\pm$ 0.096	0.838 $\pm$ 0.040	92.48 $\pm$ 13.40	55.3 $\pm$ 95.4	53.6 $\pm$ 39.6
TCN	0.555 $\pm$ 0.278	0.585 $\pm$ 0.097	0.501 $\pm$ 0.269	0.555 $\pm$ 0.278	18.75 $\pm$ 26.31	3031 $\pm$ 3125	3526 $\pm$ 2855

Table 5.2: The results of different contact classification methods, data scaling and data smoothing methods. Metrics are stated as mean and standard deviation over five cross-validation. The best value of each metric is highlighted.

Method	Scaling	Smoothing	F1	ACC	PPV	TPR	CCA	TT-S	TT-E
CPD	-	low-pass	0.858 ± 0.049	0.854 ± 0.049	0.873 ± 0.047	0.844 ± 0.052	**1.13** ± 0.04	521.6 ± 152.9	509.3 ± 189.3
RF	live	-	0.944 ± 0.017	0.944 ± 0.016	0.965 ± 0.009	0.944 ± 0.017	140.39 ± 23.97	1.2 ± 0.3	**25.2** ± 7.2
RF	lookup	-	0.947 ± 0.016	0.945 ± 0.017	0.938 ± 0.024	0.947 ± 0.016	121.33 ± 13.96	34.5 ± 42.8	49.6 ± 55.7
RF	-	low-pass	0.772 ± 0.018	0.759 ± 0.021	0.754 ± 0.026	0.772 ± 0.018	453.36 ± 24.45	21.5 ± 14.2	31.5 ± 33.2
RF	live	low-pass	0.944 ± 0.016	0.944 ± 0.016	0.965 ± 0.011	0.925 ± 0.029	142.87 ± 24.46	1.2 ± 0.3	27.4 ± 6.1
RF	lookup	low-pass	0.947 ± 0.016	0.945 ± 0.017	0.938 ± 0.024	0.957 ± 0.016	119.93 ± 14.07	34.5 ± 42.9	49.7 ± 55.4
LSTM	live	-	0.976 ± 0.008	0.975 ± 0.008	0.975 ± 0.006	0.976 ± 0.008	15.16 ± 2.68	1.2 ± 0.1	28.9 ± 5.8
LSTM	lookup	-	0.945 ± 0.014	0.943 ± 0.015	0.939 ± 0.027	0.945 ± 0.014	32.77 ± 9.74	46.5 ± 37.4	64.5 ± 60.2
LSTM	-	low-pass	0.866 ± 0.014	0.860 ± 0.017	0.864 ± 0.031	0.866 ± 0.014	74.73 ± 6.97	9.7 ± 4.7	35.8 ± 6.3
LSTM	live	low-pass	0.976 ± 0.007	0.975 ± 0.007	**0.976** ± 0.007	0.976 ± 0.016	11.16 ± 2.29	**1.0** ± 0.2	55.4 ± 48.2
LSTM	lookup	low-pass	0.947 ± 0.016	0.945 ± 0.016	0.942 ± 0.017	0.953 ± 0.027	25.03 ± 3.18	31.0 ± 42.5	87.3 ± 63.3
TCN	live	-	**0.977** ± 0.007	**0.976** ± 0.007	0.972 ± 0.008	0.977 ± 0.007	43.80 ± 12.00	4.5 ± 7.0	28.0 ± 11.2
TCN	lookup	-	0.948 ± 0.016	0.945 ± 0.017	0.936 ± 0.020	0.948 ± 0.016	74.35 ± 9.99	37.0 ± 43.7	114.1 ± 58.0
TCN	-	low-pass	0.703 ± 0.026	0.580 ± 0.092	0.584 ± 0.112	0.703 ± 0.026	12.88 ± 16.01	1842 ± 179	2262 ± 284
TCN	live	low-pass	0.976 ± 0.008	0.975 ± 0.008	0.972 ± 0.010	**0.981** ± 0.012	16.87 ± 4.43	1.4 ± 0.7	46.0 ± 27.5
TCN	lookup	low-pass	0.948 ± 0.014	0.946 ± 0.014	0.946 ± 0.026	0.951 ± 0.023	40.3 ± 10.73	37.0 ± 44.5	74.4 ± 56.7

for RF are on par. Especially with respect to the metrics CCA, TT-S and TT-E, a difference between the two scaling approaches can be seen across all methods, with *live scaling* performing best by a large margin.

This could be due to the fact that in the case of *live scaling*, each recording is considered individually. In contrast, for *lookup scaling*, the assumption is made that the same scaling value can be used for each recording with the same setting. However, since no external influences (e.g. behavior of the irrigation function, residual tissue in the instrument, etc.) are taken into account, theoretical deviations in the actual scaling value can occur, consequently providing an inadequate basis for classification.

A particularly noticeable difference can be observed in the TT-S and TT-E metrics. Independent of the scaling method, better results can be seen for TT-S across all classification methods. This means that the starting point of the *contact* area is better predicted than the endpoint. It can be assumed that this occurs because the signals have a slightly slower transition to the *no contact* level during contact release than during initial contact. This behaviour can be seen in Figure 5.2. The initial contact marked by

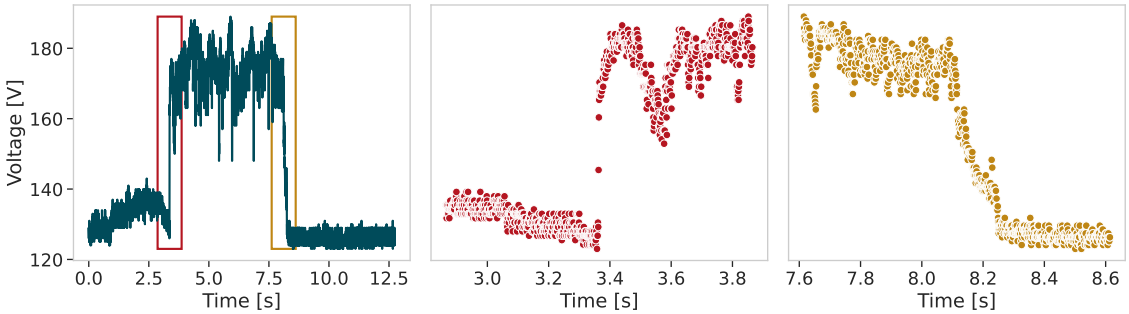


Figure 5.2: Example of the initial (red box) and final (gold box) contact area. The left image shows the entire recording, the center image shows a zoomed-in view of the initial contact, and the right image shows a zoomed-in view of the final contact. A faster transition between the *no contact* and the *contact* area can be seen in the initial contact, while the final contact shows a slower transition with multiple data points in between.

the red box on the left image and zoomed in the center image shows a faster transition between the baseline and the contact area with only a few data points in between. In contrast, the final contact marked by the gold box on the left image and zoomed in the right image shows a slower transition with multiple data points in between. One reason for this difference in behaviour could be that the oscillation of the system is damped instantaneously by the tissue contact, while on the release of the contact this is a more gradual process. This gradual process could be influenced by a time delay from residual tissue and fluids remaining within the instrument. Despite the release of contact, the aspiration process is not sufficiently rapid to eliminate the remnants. This delay persists due to limitations in the aspiration system, which is not designed to effectively address this specific scenario within a short time frame. Generally speaking,

a stronger weighting can be attributed to TT-S since immediate detection of the *contact* is more important – so that the tissue can be classified as quickly as possible in a further step – than detachment of the *contact*.

Performance Effect – Smoothing A Butterworth *low-pass* filter is applied to remove high-frequency components from the time-series data. The results of these experiments are presented in Table 5.2. Overall, the effects of data smoothing seem to depend on the classification methods at hand. For the methods CPD and RF, virtually no improvement in the set of metrics as a whole can be observed. In contrast, the deep learning-based methods LSTM and TCN achieved minor improvements. Similar to the results for data scaling, the best improvement of 0.148 in the F1-score is evidently using TCN. Thus, TCN using *low-pass* attained the performance of about 0.703 ± 0.026. Those improvements on deep learning-based approaches could be attributed to the reduction of the high-frequency components of the signals, which TCN and LSTM consequently do not have to consider. Thus, the data are shown to have less complexity, which subsequently simplifies the learning part of the networks.

Performance Effect – Scaling and Smoothing Following both preprocessing approaches, data scaling and data smoothing are combined in Table 5.2. In the case of F1, ACC, PPV and TPR metrics, no performance change through the addition of smoothing the data is recognizable. Minor deteriorations and improvements in performance can be observed in the TT-E metric, preventing a clear conclusion. However, performance improvement in the CCA and TT-S metrics can be seen for the TCN and LSTM methods. These show the highest performance, especially for LSTM, with a mean value of 11.16 ± 2.29 for CCA and 1.0 ± 0.2 ms for TT-S. Relatively speaking, metrics improve the most for TCN. At the same time, improvements on CCA and TT-S are accompanied by a reduction in TT-E performance. Nonetheless, since the weighting of importance is higher for the initial contact change and *contact* area, better CCA values and TT-S values are more relevant than TT-E values.

The Relationship of Settings and Tissue Classes to Classification Results From the previous results, a relationship between the results of the methods and the tissue samples is suspected. As a result, the performance relative to settings and tissue classes is examined more closely. Exemplary and due to the robust performance, Table 5.3 shows the calculated average F1 grouped by settings and tissue classes of the LSTM method using *live scaling*.

One can see that certain settings work better or worse with certain tissue classes than others. For example, a larger decrease in the metric for *low* setting and *firm* tissue class can be observed. This might be because a low setting encounters too much resistance when dealing with very stiff tissues, preventing a clean resection. Alternatively,

a relatively higher load on the system could produce a clearer signal, leading to improved classification results. To reach a definitive conclusion, further research is necessary.

Table 5.3: Results of LSTM using *live scaling* grouped by settings and tissue classes. F1-score is stated as mean and standard deviation.

Settings	Tissue Classes		
	soft	*medium*	*firm*
low	0.992 ± 0.007	0.967 ± 0.038	0.809 ± 0.194
moderate	0.975 ± 0.017	0.977 ± 0.018	0.991 ± 0.007
intermediate	0.994 ± 0.003	0.993 ± 0.007	0.984 ± 0.012
high	0.971 ± 0.020	0.991 ± 0.001	0.990 ± 0.004

Prediction Time For real-time application of the classification methods, the execution time of one prediction is of interest. Within the scope of prediction time analysis, prediction for CPD refers to the statistical analysis. To calculate a representative time of one prediction, the mean time of 30 predictions on different inputs is calculated and the mean of all five cross-validation models are presented in Table 5.4. Times are calculated on a Linux machine with an Intel Xeon Platinum 8168 CPU and an NVIDIA Tesla V100 GPU. The CPD is the fastest method with about 0.764 ± 0.373 ms per prediction, followed by RF with 9.111 ± 0.269 ms per prediction, LSTM with 27.755 ± 0.130 ms per prediction and then TCN with 32.497 ± 0.236 ms per prediction. This indicates feasibility for real-time application of all methods. However, with respect to the results of CPD, it must still be mentioned that a waiting time of 100 ms needs to be considered. Therefore, by assessing prediction times alongside F1 performances, it is concluded that LSTM is the best-performing method for real-time application in the context of contact detection.

Table 5.4: Prediction time of all four methods. Time is provided as mean and standard deviation.

Method	Prediction time [ms]
CPD	0.764 ± 0.373
RF	9.111 ± 0.269
LSTM	27.755 ± 0.130
TCN	32.497 ± 0.236

5.3 Feasibility under Realistic Conditions

As previously shown, data acquired under controlled conditions inhibit enough information to show the feasibility of classifying when contact is established between the

ultrasonic aspirator and tissue. However, this does not reflect a real-world scenario. In a clinical setting, the usage of an ultrasonic aspirator is solely conducted in a handheld manner. This of course introduces additional unknown variables such as the surgeon's hand tremor, speed of movement, the force applied to the tissue and angle of the instrument. Those variables are always present and can only be partially controlled and make the contact classification more challenging than in a controlled environment.

Data Preparation To investigate the feasibility of the classification methods under realistic conditions, the data set D6 is used for training purposes while the data set D7 is used for testing given a highly diverse set of data recorded under different force conditions. Due to the sheer size of the training data set of more than 82 million data points, a manual labeling of contact ground truth is not feasible. As already mentioned in the data chapter 4.2.1, an offline-based contact classification was therefore carried out. This classification is however only given as a rough estimation and cannot be guaranteed to be 100 % accurate for all data points. As is well known in the field of machine learning, the use of a suboptimal data set can result in suboptimal outcomes. Consequently, the data set is subjected to a rigorous process of processing and cleansing. For this purpose data that is considered as a steady state, i.e. regions of the data in which no change between *contact* and *no contact* is present, is omitted. This is done since data in this area could be given false negative labels if the automatic classification missed a contact, which could teach the model wrong information and hinder the learning process severely. Therefore, only data in which a change between *contact* and *no contact* is occurring, as well as an area of half a second before and after this point, is considered for training. This approach enables the representation of data points that reflect a change, while still maintaining the individual states on either side of the change due to the half-second window. Thus, all possible states are represented in the training data set and can be learned with an adequate respective field of view of the model. However, the data set is further refined to ensure high-quality data, with the overarching goal of identifying only reliable contacts. By assuming that a contact has a higher voltage level during *contact* than during *no contact*, while remaining relatively stable within these states, we filter out data points that do not meet this criterion. This process helps in eliminating inaccurate data points with incorrect ground truth labels, allowing the model to learn the correct features from the data and focus on identifying reliable contacts. Data are scaled similar to using the *live scaling* approach, but with the difference that the scaling values are used for dividing the data points instead of subtracting them. This way, the data are scaled to a relative range instead of absolute values with the *no contact* values around one. This allows the data to retain its relative shift in relation to the base level, which can deliver additional information in the context of more complex data, specifically with the new unknown variables introduced in the handheld scenario. The scaling value is based on the mean of the first 50 data points

of the continuous data from a one-second window centered around the change point going from *no contact* to *contact*. The accompanied one-second window of the reverse change is using the same scaling value. In other words, each individual contact and the associated contact loss are scaled to the same level. It is important to note that this process is repeated for each contact, regardless of whether they originate from the same recording sequence.

In contrast to the previous data set for training, the test data set D7 is not labeled automatically but manually due to its size. Labeling is carried out by using the open source software Label Studio [174]. This methodology allows for the data to be labeled with a high degree of accuracy, while maintaining the integrity of the data set and preventing any omissions. This approach thus yields an optimal data set for testing the classification methods under realistic conditions. As the data sequences are unsegmented and are intended to represent actual live data, the data are scaled in the same manner as the training data set, but for the entire sequence rather than for each individual contact.

Both data sets are resampled to 350 Hz to alleviate computational burden while retaining the ability to promptly detect changes in contact state in real-time. This study only relies on the single voltage parameter for classification, as the manual labeling process is limited to this parameter and provides enough information for the task at hand for a human observer.

Methods Three different classification methods are used to investigate the feasibility of the classification under realistic conditions and are based on the findings of the previous experiments under controlled conditions. Initially, a baseline using a simple threshold-based approach is used which does not include any time-based information – an even more simplified version of the previously used RF method. This approach is based on the assumption that the voltage level during *contact* is higher than during *no contact*. Thus, the voltage distribution of the training data set is analyzed. An optimal threshold is chosen by computing the receiver operating characteristic (ROC) curve between class labels and voltage and selecting the threshold that maximizes the true positive rate subtracted by the false positive rate. This optimal threshold is then used for the classification of the test data set and is referred to as *ROCTh*.

The second approach combines the threshold-based approach with a calibration step as well as an intital contact detection step. An initial contact detection is based on the sudden change in voltage level. The equation for this is given in the following:

$$\bar{V}_t^{\text{prev}} = \frac{1}{w} \sum_{i=t-w-1}^{t-1} V_i \tag{5.1}$$

$$value_t = \frac{(V_t - \bar{V}_t^{\text{prev}})^2}{\bar{V}_t^{\text{prev}}} \cdot \text{sign}(V_t - \bar{V}_t^{\text{prev}}). \tag{5.2}$$

w is the window size in which the last w data points are considered for the mean calculation of the previous voltage level $\bar{V}_t^{\text{prev}}$ at time t. The current voltage value V_t is then compared to the previous voltage level to calculate the value $value_t$ by squaring its difference and dividing it by the previous voltage level and multiplying it by the sign of the difference. $value_t$ is calculated for each data point of the training data set in which an initial contact is present. Due to the training data setup, each one-second windowed contact has the change of contact in the center. Thus, the maximum $value_t$ around the change point is used (interval of 25) for each change from *no contact* to *contact*. The fifth percentile of those aggregated values is then used as the threshold for classifying the change from *no contact* to *contact*, similar to the CPD method. The reverse change from *contact* to *no contact* is not considered in this step due to slow changes in the voltage level in this case which is shown in Figure 5.2 for controlled conditions using a CNC machine, which still applies to realistic handheld conditions. Therefore, detecting when the contact is lost is solved by a different approach. If the voltage level drops below the current voltage level of *contact* and reaches the voltage level of *no contact*, the contact is considered lost. However, this can be difficult if a shift of voltage level over time is not considered as shown in Figure 5.3. Ideally, the voltage level of *no contact* should

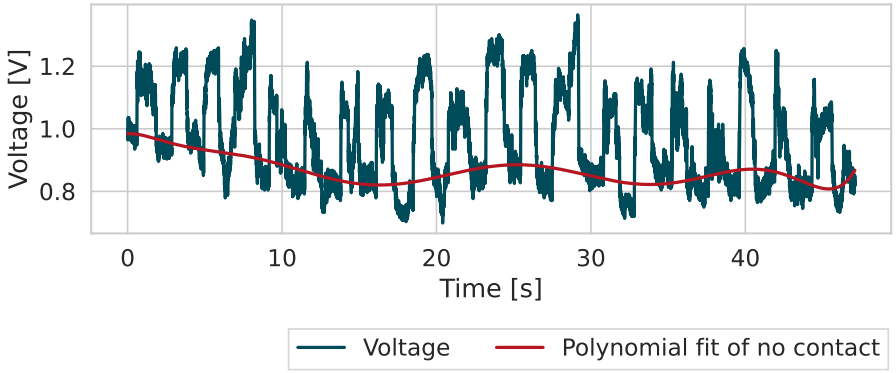


Figure 5.3: Voltage drift over time on test data sequence. The red line shows the polynomial fit of the voltage level of *no contact* over time for better visualization.

stay constant over time. As shown, this is not always the case and the voltage level drifts over time. The red line shows the polynomial fit of the voltage level of *no contact* over time for better visualization. Due to complex interactions between the ultrasonic asprirator's sonotrode and fluid dynamics of its irrigation system as well as tissue waste of its aspiration system, which can be left in the instrument, the voltage level might be prone to changes over time. For this reason, a dynamic threshold is calculated for detecting when the contact is lost and reaching the voltage level of *no contact* again.

This dynamic threshold is calculated by the following equations:

$$d = \frac{1}{100}\sum_{i=1}^{100} \mathbf{b}_i^{\mathrm{c}} - \frac{1}{100}\sum_{i=1}^{100} \mathbf{b}_i^{\mathrm{nc}} \tag{5.3}$$

$$dyn\_threshold = \frac{1}{100}\sum_{i=1}^{100} \mathbf{b}_i^{\mathrm{nc}} + \frac{|d|}{4}. \tag{5.4}$$

The absolute difference d between the mean of 100 data points i of two buffers $\mathbf{b}^{\mathrm{c}}$ and $\mathbf{b}^{\mathrm{nc}}$ of the voltage level of *contact* and *no contact* is calculated. Those two buffers are filled when a change in state is detected by the initial contact detection. The buffer $\mathbf{b}^{\mathrm{c}}$ is filled with the voltage right after the change from *no contact* to *contact* and the buffer $\mathbf{b}^{\mathrm{nc}}$ is filled with the voltage values right before this change. A zone of ten points in each direction of the change point is not considered for filling the buffers to avoid the influence of the change point. The dynamic threshold is then calculated by adding a quarter of the absolute difference to the mean of the buffer $\mathbf{b}^{\mathrm{nc}}$. Thus, if after the initial contact detection the voltage level drops below the calibrated threshold, the contact is considered lost. After the contact is lost, the initial contact detection is repeated to detect the next contact and all buffers and the dynamic threshold are reset. This approach is referred to as *DynTh* and is depicted in its entirety in the appendix in algorithm 1, allowing a classification in an online manner. The assumption for this algorithm is that the initial values w always belong to the *no contact* state, which is a valid assumption if the device is turned on before the contact is established. An additional assumption is that right after a contact change is detected, the state is stable for at least 100 data points, which is short enough to be considered a stable state considering the sampling rate of 350 Hz. The same is done for the labels before the state change. The input to the algorithm consists of a test voltage sequence and the threshold for the initial contact detection. The output is a sequence of binary values indicating the contact state at each time point as zero for *no contact* and one for *contact*.

A third approach is based on an RNN method, which has shown the best performance in the previous experiments. Specifically, an LSTM-based RNN is used for this task which uses two channels of a sequence of 100 data points as input and predicts all data points in the 100 data point sequence as either *contact* or *no contact*. The first channel represents a voltage sequence, whereas the second channel incorporates a differentiated signal, which is obtained by subtracting the value of the voltage at a given instant from the value of the voltage at the previous instant. This process captures the rate of change of the voltage signal. The RNN is trained on the training data set with a reduced architecture of two LSTM layers each with four nodes and a single linear output layer followed by a sigmoid activation function. An RAdam optimizer [108] is used with a binary cross-entropy loss function and an initial learning rate of 0.001. The model is trained for 100 epochs with a batch size of 128 and a learning rate scheduler which

reduces the learning rate by a factor of 0.1 if the validation loss does not improve for 20 epochs. If no improvement is seen for ten epochs, the training is stopped. Additionally, a dropout of 0.2 is used for the LSTM layers to prevent overfitting. The already processed training data are normalized by subtracting the mean and dividing by the standard deviation over all data points. This calculated mean and standard deviation is then used for the normalization of the test data set. Since the training data are initially scaled to having the *no contact* state around one (leading to values of lower than zero after normalization), the previous mentioned shifting of voltage in a test sequence can lead to suboptimal results. To overcome this, a dynamic recalibration of the test data set is performed during online inference in a similar fashion to the *DynTh* method. The idea is to shift the voltage data to reflect the distribution of the training data set after normalization. The distribution is shown in Figure 5.4. It can be seen that the

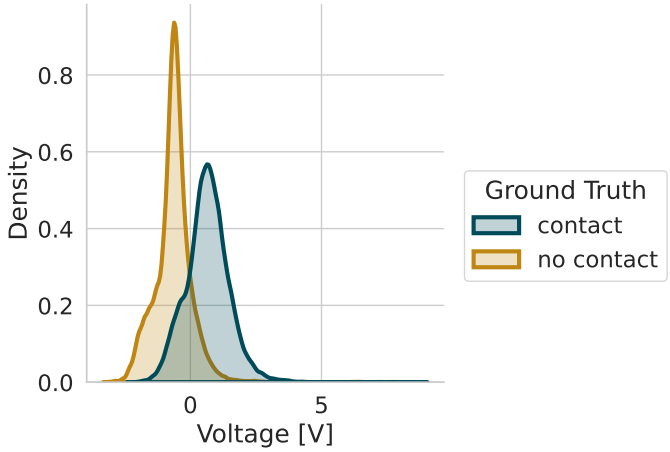


Figure 5.4: Voltage distribution of training data set for LSTM method.

contact state mean is larger than zero and the *no contact* state mean is smaller than zero; consequently, the center between the two states is around zero. The recalibration value is computed using the following equation:

$$cal\_value = \frac{1}{100} \sum_{i=1}^{100} \mathbf{b}_i^{\mathrm{nc}} + \frac{d}{2}. \tag{5.5}$$

The recalibration value, denoted as cal_value, is determined by adding half of the difference d between the means of the two buffers $\mathbf{b}^{\mathrm{c}}$ and $\mathbf{b}^{\mathrm{nc}}$ to the mean of buffer $\mathbf{b}^{\mathrm{nc}}$. Essentially, this yields the midpoint between the buffer means, thus indicating its deviation from zero. The cal_value is subsequently employed to adjust the test data distribution towards that of the training data set by subtracting it from the voltage data, thereby shifting the center closer to zero. The entire process is depicted in the appendix in algorithm 2. Again, the first 100 initial values are assumed to belong to

the *no contact* state. The $cal\_value = 0$ is set at the beginning of the algorithm and is recalculated on every new encounter of a state change from *no contact* to *contact*. That is if the difference of the means of the buffers $\mathbf{b}^{\mathrm{c}}$ and $\mathbf{b}^{\mathrm{nc}}$ is above 0.3 to ensure that the recalibration is only done if the data are in a stable state. Otherwise, the previous cal_value is used. In a way, this approach is detrending the data during inference to ensure that the model is fed with data that is similar to the training data set. This process transfers the condition of the training data, which was rescaled for each contact, to the test data.

As in the previous experiments, similar metrics such as F1, PPV and TPR are used to evaluate the performance of the methods. Due to the majority of the data being in the *no contact* state, the F1, PPV, and TPR metrics are weighted, while instead of the ACC a balanced accuracy (BA) is used.

These metrics provide a comprehensive overview of performance across the entire data set. The metric *counted contact areas* CCA cannot be utilized here, as the test data consists of complete sequences rather than individual contacts. Thus, the amount of contacts per sequence differ, making the CCA metric unsuitable for this evaluation. To accommodate test data comprising sequences lasting up to 3.8 min with multiple contacts, the evaluations focus on specific areas of interest: the start and end of each contact. These areas are defined around the change points in contact states, encompassing half a second before and after each change. Within each start and end area, previously utilized TT-E and TT-S metrics measure the deviation between predicted and actual change points in milliseconds. Results are organized by contact start and end, abbreviated as TT with suffixes denoting the area of interest (-S for start and -E for end when necessary). The MSS metric evaluates the percentage of missed contacts and missed contact ends. Additionally, accuracy within these areas is assessed using the ACC metric to provide a more comprehensive understanding of performance. Both the MSS and ACC are appended with -S for the start of a contact and -E for the end of a contact if deemed necessary in the text.

Various experiments are being conducted to evaluate the performance of different methods under realistic conditions. Initially, the overall performance of these methods is assessed using a designated test data set. Subsequently, variations in training procedures on generator setting dependences and their effects on performance outcomes are examined. Furthermore, an experiment is conducted to investigate the potential performance enhancements achieved by employing an ensemble approach comprising multiple similar methods. The need for recalibration is also examined to determine its impact. Finally, the relationship between the force applied to the tissue and the classification performance of the methods is analyzed to measure the influence of force on the performance metrics.

General Results on Test Data The optimal threshold for the *ROCTh* method is determined to be 1.102, which is applied in classifying the test data set. In contrast, the *DynTh* method utilizes the fifth percentile of the $value_t$ calculated from the training data set, resulting in a threshold of 0.166. This translates to a transition from *no contact* to *contact* when the current value exceeds the previous mean value by 16.6 %. The evaluation of the three approaches – *ROCTh*, *DynTh*, and *DynRNN* – is conducted on the test data set D7, with the results presented in Table 5.5. It can be seen that out of the three methods,

Table 5.5: Results of the three methods on the test data set for contact detection under realistic conditions. The best value of each metric is highlighted. Metrics for start and end of contact are explicitly stated.

Method	BA	F1	PPV	TPR	Start			End		
					TT	ACC	MSS	TT	ACC	MSS
ROCTh	0.853	0.868	0.868	0.869	36.2 ± 46.8	0.861	6.4	64.8 ± 73.4	0.760	15.5
DynTh	**0.880**	**0.891**	**0.891**	**0.892**	**27.4** ± 39.4	0.886	15.7	69.7 ± 76.2	0.783	27.7
DynRNN	0.875	0.882	0.882	0.882	28.0 ± 41.7	**0.921**	**1.1**	**54.5** ± 57.1	**0.813**	**2.4**

ROCTh performs suboptimally in comparison to the other two methods, which is to be expected due to the simplicity of the method using only a single threshold for all data points and no time-based information. When taking a look at the overall metrics BA, F1, PPV and TPR, the *DynTh* method performs best with a BA of 0.880, an F1 of 0.891, a PPV of 0.891 and a TPR of 0.892, followed closely by the *dynRNN* method.

Since specific areas such as the start and the end of a contact are of interest, the metrics within half a second before and after the change points are evaluated. Very similar to the previous study on controlled data, a difference between the two metrics TT-S and TT-E is observable. All methods struggle with the detection of the end of a contact, which is shown to be twice the value of TT-S. This is due to the nature of the data, where the end of a contact is not as abrupt as the start of a contact (c.f. Figure 5.2). Thus, making a clear distinction between the two states is more difficult in this online classification approach. The difference in transition change between the two states could be due to the sudden coupling of mass on the system at the beginning of a contact, which is not present at the end of a contact. Moreover, at the end of a contact, the instrument is carrying tissue debris and fluids of the resection, which can lead to a more gradual decrease in voltage. Reaching a definitive conclusion about this phenomenon is proving difficult, emphasizing its complexity and the need for further research using specialized methodologies to unravel its intricacies. This difference is also reflected in the remaining metrics. Even though the *DynTh* method performs best in the overall metrics, the *DynRNN* method shows the best performance in the areas of interest with an ACC-S of 0.921 and only 1.1 % missed starts of a contact in MSS-S. Similarly, the highest performance in the end of a contact is shown by the *DynRNN* method with an ACC-E of 0.813 and only 2.4 % missed ends of a contact in MSS-E. Interestingly, a similar TT-S for

both the *DynTh* and *DynRNN* methods is observed, which on the other hand is not the case for the ACC-S metric.

As previously mentioned, a higher importance is given to the start of a contact as an immediate response is needed for subsequent actions as tissue classification. A missed or false detection of losing contact is considered less critical, as the instrument would no longer be in contact with the tissue or only communicating contact loss, thus causing no harm. The difference between the two metrics is caused by the definition of the TT-S metric which only shows the closest data point to the change point, while the ACC-S metric measures the entire area around the change point. This indicates that more change points – false negatives and positives – further away from the actual change point are detected by the *DynTH* method, which is not the case for the *DynRNN* method. Meaning that the *DynRNN* method is more reliable in detecting the actual change point within the vicinity of the change and reduces the number of false positives and negatives, facilitating a more accurate classification of this area.

Setting Dependency With the ultrasonic aspirator's generator settings being a crucial factor in the resection process, the relationship between the settings and the contact classification results is investigated. This is done by evaluating the performance of the results from Table 5.5 grouped by the generator settings *low*, *medium*, and *high*. Due to the results only being evaluated and not trained depending on the settings of the generator, they are not setting-dependent. In order to evaluate the performance under the dependency of the generator settings, all methods are trained on their setting-dependent data set and are evaluated on the according test data set. Thus, each setting has its own trained model instead of a single model for all settings. A visual representation of the results is shown in Figure 5.5 for the metrics F1, ACC-S and MSS-S. Quantitative results are given in Table B.1.

For both the *ROCTh* and the *DynTh* methods, training depending on the setting leads to a slight improvement or at least on par performance in the metrics. This is especially noticeable in the MSS-S metric which results in fewer missed starts of a contact. This, however, is further dependent on the selected setting. It can be seen that throughout, both the *ROCTh* and *DynTh* methods, with and without setting dependence, a similar pattern exists with the *medium* setting showing a slightly stronger performance in the metrics F1 and ACC-S and a more pronounced difference in the MSS-S metric. Even more pronounced is the difference in the *DynRNN* method. It can be seen clearly that a difference between the means of a setting-dependent and setting-independent training is present with the setting independent training showing a better performance in all metrics. Especially noticeable is the lack of missing contact start for two out of three settings in the MSS-S metric, showing a strong performance in the start of a contact detection which is similarly present in the ACC-S metric. The difference in between the setting-dependent and independent training is most likely due to the amount

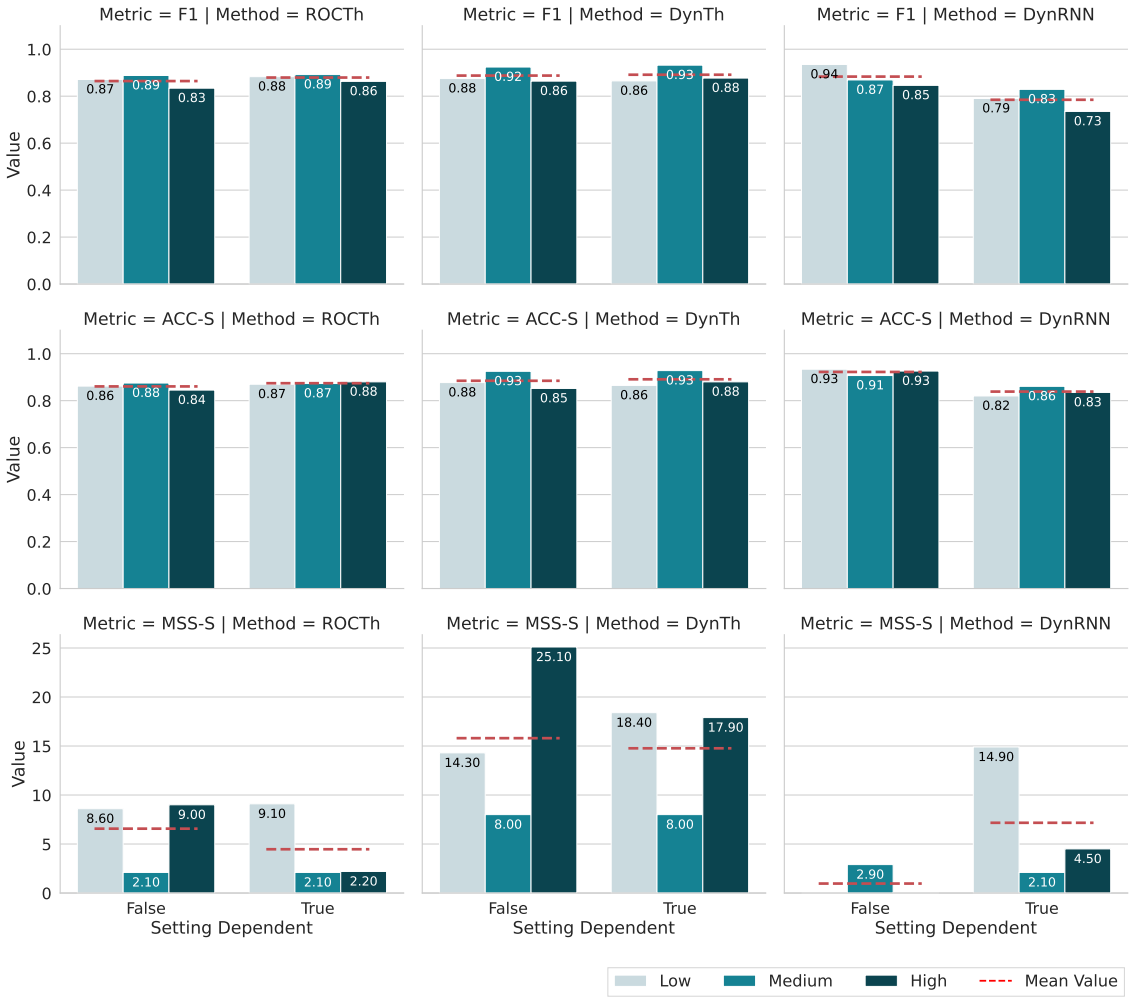


Figure 5.5: Results on test data, grouped by setting, with and without setting dependence. Setting-dependent indicates a model for each setting, rather than one model for all settings. Dashed red line shows mean per setting dependence.

and variety of data present in the training data set which is obviously smaller in the setting-dependent data set due to it being a subset of the setting-independent data set. Thus, this leads to a more robust model which can generalize better to unseen data. With a larger amount of data, a slightly increased performance might be observed in a setting-dependent training, as indicated by the other methods. This however, is out of the scope of this study and should be investigated in future studies.

Ensemble Method Given the *DynRNN* method's performance as a standalone method, the possibility of combining multiple instances as an ensemble framework is investigated. An ensemble could potentially improve the performance of the method by

combining the predictions of multiple instances, thus reinforcing the strengths of the individual models while mitigating their weaknesses. Due to the *DynRNN* method's dropout function, each trained instance is different from the others, allowing for a diverse ensemble. An ensemble of five instances is created by training five different models with the same architecture and hyperparameters. The predictions of the five models are then combined by averaging the output of each model per data point. This gives a value for each data point between zero and one. A decision threshold is set to being greater than 0.4, meaning that if at least three out of the five models predict *contact*, the ensemble predicts *contact* as well. The results – individually for each ensemble member and the combined ensemble – are given in Table 5.6.

Table 5.6: Results of the ensemble method on the test data set for contact detection. The best value of each metric is highlighted. Metrics for start and end of contact are explicitly stated.

Ensemble Member	BA	F1	PPV	TPR	Start			End		
					TT	ACC	MSS	TT	ACC	MSS
1	0.882	0.882	0.882	0.882	28.0 ± 41.7	0.921	1.1	54.5 ± 57.1	0.813	2.4
2	0.813	0.813	0.817	0.813	30.2 ± 45.4	0.910	2.8	60.4 ± 64.5	0.775	5.6
3	0.873	0.890	0.892	0.892	**27.1** ± 38.2	0.925	0.5	61.4 ± 64.5	0.794	6.1
4	0.888	0.898	0.897	0.898	27.1 ± 40.2	0.933	**0.2**	60.8 ± 68.0	0.826	5.8
5	0.877	0.883	0.883	0.883	28.8 ± 42.0	0.914	0.6	**49.1** ± 52.5	**0.830**	2.0
Ensemble	**0.892**	**0.902**	**0.902**	**0.903**	27.1 ± 39.7	**0.934**	0.3	52.5 ± 58.5	0.829	3.3

The results highlight the potential of the ensemble method, with the ensemble outperforming all individual models in the metrics considering the entire sequences like BA, F1, PPV, and TPR. In the two specific areas of interest – the start and end of a contact – the ensemble method shows a similar performance to their best individual model of the ensemble. This indicates that especially in areas outside of interest, meaning solely *contact* and *no contact* states, the ensemble method can improve sections in which individual models struggle and thus improve the overall performance of the method. Within the areas of interest, an ensemble method is able to maintain the performance of the best individual model, giving more confidence in the overall performance of the deep learning-based method.

To give a better understanding of the performance, the best two and worst two sequences of the ensemble method regarding their F1-score are shown in Figure 5.6. The best sequences have an F1-score of 0.98 and 0.96, while the worst sequences have an F1-score of 0.77 and 0.79. It can be clearly seen that the predictions of best sequences are very close to the underlying voltage signal, with the occasional missed contact during a contact state. However, if one considers the temporal context of the lost contacts, it becomes evident why a *no contact* state is predicted here. For example, in the top signal of the second and third contact, a sudden drop can be observed during the contact. From the perspective of the *DynRNN*, which only processes the data in a stream

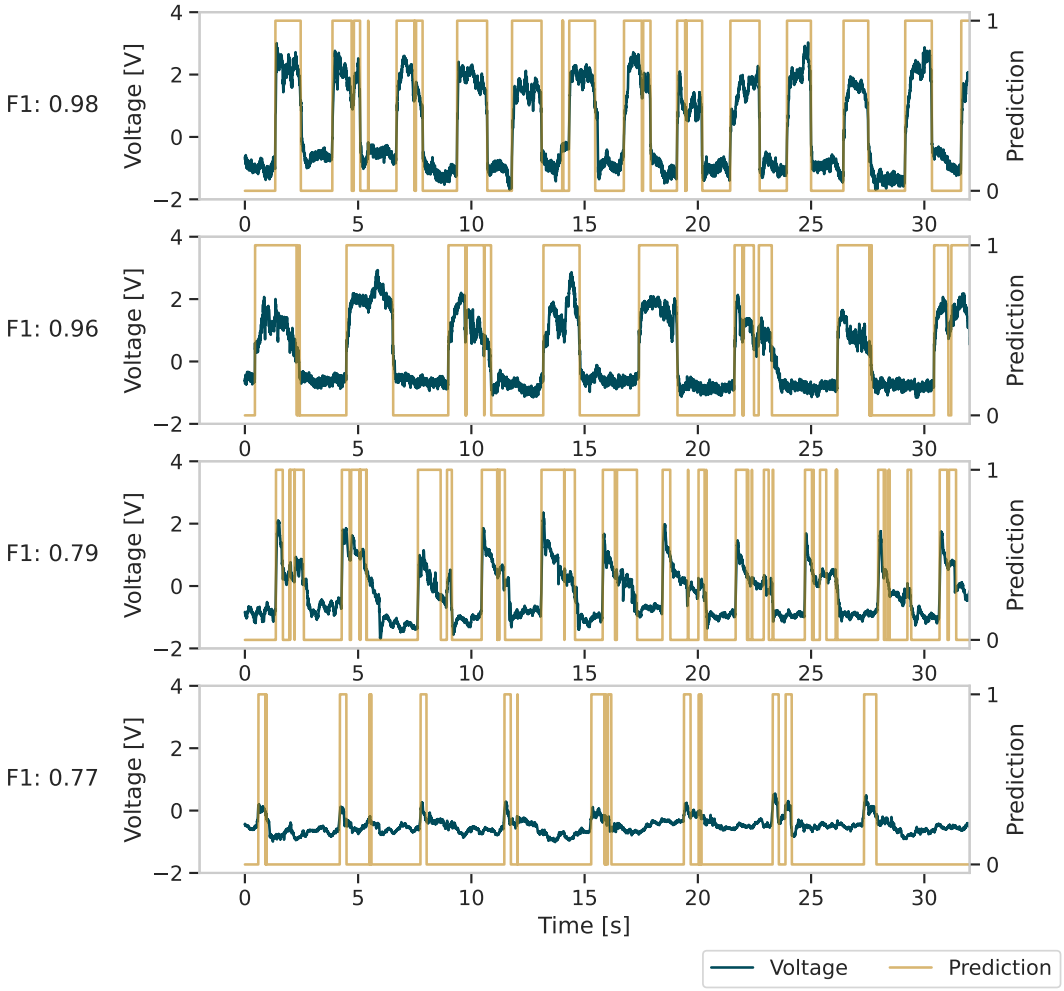


Figure 5.6: Best and worst sequences of the ensemble method regarding their F1-score. The top two rows show the best sequences, while the bottom two rows show the worst sequences – limited to first 30 s of the sequence for better visibility. The blue voltage signal is on the left, and the gold predicted contact state (0 for *no contact*, 1 for *contact*) is on the right. Voltage signal is standardized by subtracting the mean and dividing by the standard deviation based on the training data set and displayed as not calibrated.

of incoming data points, it logically interprets a loss of contact when it falls, given its limited information. In the broader context, it is evident that no total loss of contact has occurred. However, it is important to note that it is not possible to definitively ascertain what occurs at such points. It is possible that a loss occurs on a microscopic level, as the tissue is continuously removed or that the user is moving the instrument upwards for a short period of time. This cannot be answered within the scope of this work and must be evaluated here as *contact*.

A comparable phenomenon of contact loss can be observed in the second extreme case (F1 of 0.79). In this instance, the initial contact value exhibits a pronounced overshoot, which subsequently declines rapidly and is not conducive to reliable classification. In the most unfavorable prediction scenario with F1 of 0.77, it is evident that the signal is approaching the threshold of noise. Nevertheless, the *DynRNN* is capable of discerning the initial point. In the case of such signals, it may be beneficial to consider introducing an additional quality that only allows a prediction under certain conditions. This could help to ensure the reliability of potential downstream tasks, such as tissue classification.

Necessity of Calibration The effect of recalibration on online classification on the test data set is evaluated. This is done by comparing the results of the *DynRNN* method with and without recalibration. No recalibration is performed by only utilizing the LSTM models prediction without using algorithm 2. In both cases the results are acquired using an ensemble of five models each and are presented in Table 5.7.

Table 5.7: Results of using the *DynRNN* method with and without recalibration on the test data set. The best value of each metric is highlighted. Metrics for start and end of contact are explicitly stated.

Recalibration	BA	F1	PPV	TPR	Start			End		
					TT	ACC	MSS	TT	ACC	MSS
No	0.863	0.877	0.877	0.878	30.8 $\pm$ 41.4	0.909	1.2	**48.5** $\pm$ 55.1	0.807	6.2
Yes	**0.892**	**0.902**	**0.902**	**0.903**	**27.1** $\pm$ 39.7	**0.934**	**0.3**	52.5 $\pm$ 58.5	**0.829**	**3.3**

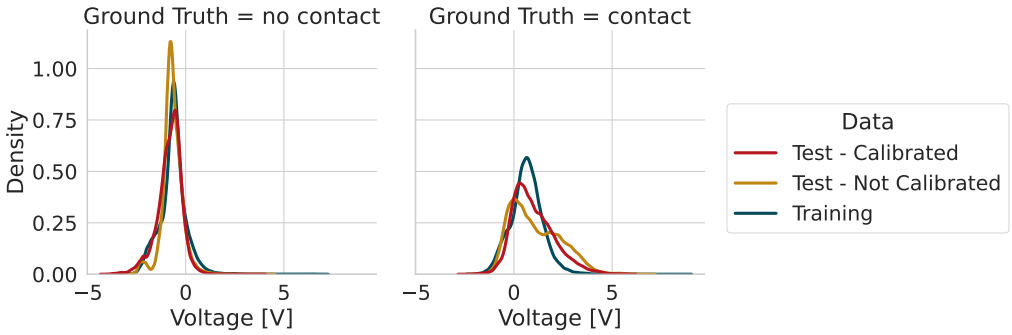


Figure 5.7: Voltage distribution on the test data set with and without the recalibration method applied, as well as the training data set. A clear shift in the test data distribution is visible when recalibration is applied, aligning it closer to the training data distribution.

The impact of recalibration on the performance of the *DynRNN* method across all metrics is notable ($p < 0.05$), suggesting its necessity for online contact state classification. This necessity becomes evident when observing the voltage data distribution in Figure 5.7, where recalibration aligns the test data distribution closer to that of the training data. Without recalibration, this alignment is absent, resulting in a less effective

method. While perfect alignment between training and test data remains elusive, future efforts may improve this. Nonetheless, these findings underscore the effectiveness of the recalibration approach.

Impact of Force on Prediction Since the test data are acquired under different forces, the impact of force on the prediction is evaluated. This is done by grouping the test data set on the average force applied and evaluating the F1 performance of the *DynRNN* method as an ensemble on each sequence. The results are shown in Figure 5.8. As

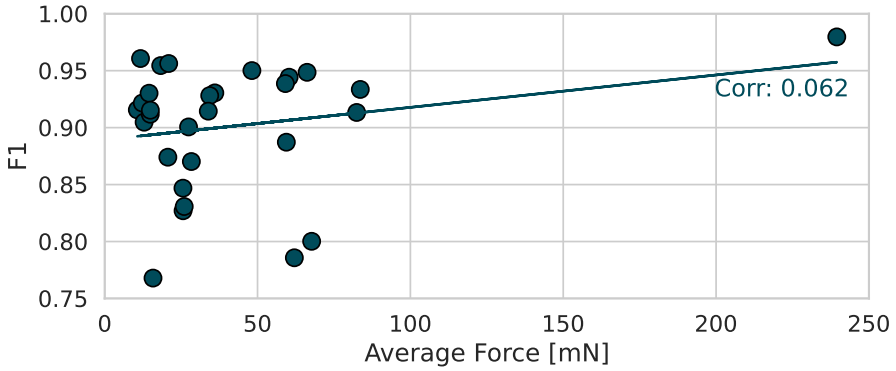


Figure 5.8: Impact of force on the F1 performance of the *DynRNN* method as an ensemble on the test data set for contact detection under realistic conditions.

depicted, no clear correlation between the average force applied on the sequence and its F1 performance is evident with a Pearson correlation coefficient of 0.062. Thus, a direct impact of force on the prediction is not observable and cannot be determined from the data.

5.4 Summary

Determining the contact state of the ultrasonic aspirator is a crucial task in the overarching goal of developing an intelligent ultrasonic aspirator. Knowing if the instrument is in contact with tissue or not is essential for subsequent tasks such as tissue classification. This would facilitate feature generation for tissue classification to a fixed point in time aligned with the moment of contact, thus enabling a reproducible setting for the classification. Furthermore, the intelligent ultrasonic aspirator process pipeline is viewed in hierarchical order, with the contact state classification being the first step as shown in Figure 5.1. Thus, subsequent tasks can only be performed if the contact state is known to be *contact*. Otherwise, a loop is created until a contact is encountered. By that, a reduction of complexity and computational resources can be achieved. Aside from the subsequent task being the goal of tissue distinction, different potential applications can be envisioned which are dependent on the contact state. For example, control of individual system components such as the irrigation and aspiration system or the ultra-

sonic power can be adjusted based on the contact state. This could lead to reduction of unnecessary energy consumption and fluid usage, thus increasing the efficiency of the system and reducing the overall cost and environmental impact. Moreover, ultrasonic power contact state dependence may result in a reduction of accidental damage to non-patient tissue, such as the surgeons themselves or other surrounding objects.

The work presented in this chapter demonstrated that contact detection on the basis of an ultrasonic aspirator's electrical signal is feasible and can be achieved with high accuracy. In a first step, contact detection was carried out within a controlled environment on artificial tissue models, where the instrument was controlled using a CNC machine. Different methods were evaluated, with the LSTM-based RNN method showing the best performance with an F1-score of about 0.97 and showcasing the importance of the temporal context in the classification as well as the necessity of optimal scaling of the data. All methods were built and evaluated in an online manner, meaning that the classification was done on a stream of incoming data points. This is crucial for the application in the intelligent ultrasonic aspirator, as the classification has to be done in real-time. Thus, prediction time was factored into the evaluation of the methods and showed manageable times of less than 30 ms for the LSTM method.

In a second step, contact detection was evaluated under realistic conditions, where the instrument was controlled by a human operator in a handheld manner on artificial tissue models. This represents the actual use case of an ultrasonic aspirator in a surgical setting. The LSTM-based RNN method was further evaluated and showed the best performance with an F1-score of about 0.90 – lower than in the controlled environment, but still a high performance given the increased complexity and uncontrollable factors of the data. Since no more complex methods were evaluated, prediction times are still assumed to be manageable for the LSTM method for real-time classification. Key to the success of the method was two-fold: the recalibration of the incoming data to align it with the training data and the ensemble method, which combined multiple models to improve the overall performance. The use of an ensemble of five models permitted a more robust and reliable classification, as the collective advantages of the individual models were combined to offset their respective weaknesses. Nonetheless, the number of models in the ensemble was not optimized and could be refined to enhance performance. Recalibration was shown to be significant for online classification, as it ensured the alignment of the test data distribution with that of the training data on which the model was trained. This approach permitted the maintenance of performance over longer test data sequences, which were susceptible to signal variations. Influences of forces and generator settings on the classification were evaluated, with no clear correlation found on the force and a slight improvement in the performance of the methods when trained depending on the generator settings in case of non-deep learning methods.

Generally, independent of investigated methods and conditions, the start of a contact was easier to detect than the end of a contact. This is due to the nature of the data, where the start of a contact is more abrupt than the end of a contact. This, however,

is not a problem since a missed end of a contact is less critical than a missed start of a contact, which builds the basis for the subsequent tasks and thus is given a higher importance.

The findings from this work could be valuable for potential feature development, particularly through the application of sensor fusion methods. For instance, the integration of microphones could offer additional insights into the current contact state, known to undergo acoustic changes when in contact with tissue, a factor not explored in this work. The processing of such acoustic signals in conjunction with electrical signals could lead to a more comprehensive understanding of the contact state.

The outcomes of this endeavor present promising advancements, demonstrating the practicality of real-time online contact detection on an ultrasonic aspirator. The integration of contact state classification with tissue resection tools represents a significant and novel advancement, particularly in the context of very soft tissue handling such as brain tissue. These results lay a strong foundation for the development of an intelligent ultrasonic aspirator, where understanding the contact state serves as the crucial initial stage towards intraoperative tissue classification in neurosurgery.

CHAPTER 6

Tissue Differentiation

Tissue differentiation with an ultrasonic aspirator is a challenging and novel task. This is a complex process, as the ultrasonic aspirator is not designed to function as a tactile sensor, despite its similar operational characteristics. Additionally, it is important not to overly restrict the primary function of the ultrasonic aspirator, which is tissue ablation, by implementing intelligent system components. The objective of this chapter is to examine the possibility of distinguishing between various tissue stiffnesses using electrical data obtained from an ultrasonic aspirator to answer research question II) introduced in chapter 1. This part is emphasized in blue in the flowchart in Figure 6.1, indicating where it fits into the overall concept of an intelligent ultrasonic aspirator. As a first step, the feasibility of using electrical data from an ultrasonic aspirator is investigated under controlled conditions, to determine if a relationship between the electrical data and the tissue stiffness can be found. In the following steps the difficulty of the task is increased step by step by including tissue ablation and handheld mode to further mimic real surgical scenarios. The next phase explores the influence of forces on tissue differentiation and feature generation, as well as the relevance of the results in a clinical context. Finally, the performance of the classifier is compared to human performance.

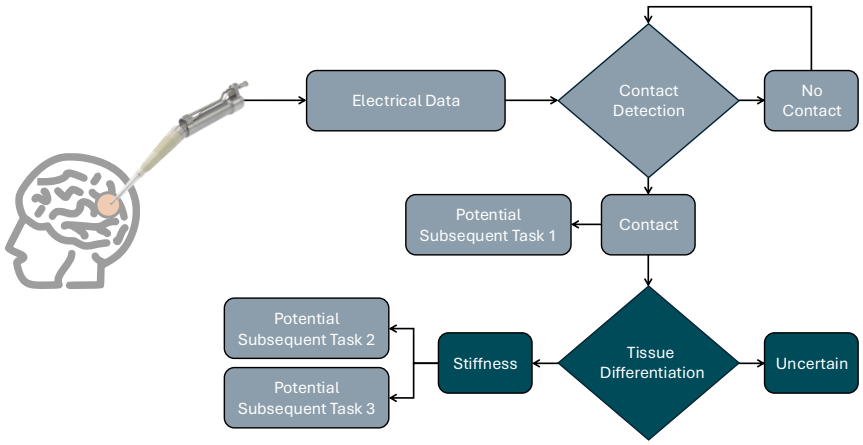


Figure 6.1: Flow chart of an intelligent ultrasonic aspirator. The content of this chapter is highlighted in blue.

6.1 Related Work

A number of studies are currently investigating the potential of tactile sensors for differentiating between various types of tissue, with a particular focus on brain tissue. In a study conducted by Tanaka et al. [170], a sensor system is employed, utilizing a fluid-filled balloon as its foundation. This balloon is inflated with a fluid, and upon mechanically interacting with the tissue, it facilitates the assessment of tissue properties. The system's efficacy was visually assessed across different scenarios, including urethane gel, as well as white and gray matter of the porcine brain. In their study, Johannsmann et al. [85] explored the potential of a piezoelectric tactile sensor equipped with torsional resonators for differentiating between different types of tissues. They focused on analyzing changes in resonance frequency and bandwidth of the sensor when transitioning from a *contact* to a *no contact* state. These changes were then linked to the distinction between healthy and tumorous brain tissue, carried out through in vivo rat experiments. The analysis revealed correlation coefficients reaching up to 0.87. Likewise, Stroop et al. [165] employed piezoelectric bimorph sensors to differentiate between various regions of ex vivo porcine brain tissue, including the cortex surface, white matter, basal ganglia, and thalamus. They achieved this by exciting the tactile sensor using multisine signals to obtain frequency response functions. Subsequently, these response functions were processed and organized into clusters using a k-means algorithm. The results showed notable discriminative ability for gray matter and thalamus tissue, whereas white matter and basal ganglia tissue exhibited comparatively reduced discriminative power. Brain tissue differentiation with stereoelectroencephalographic (SEEG) data is conducted by Machado et al. [112]. In some cases of drug-resistant epilepsy patients, SEEG electrodes

are implanted in the brain for diagnostic purposes. The authors used pairs of SEEG electrodes to classify whether the tissue of implantation is gray or white matter. Feature extraction is performed on Bode plots of the voltage SEEG signals, followed by a classification with a Linear Discriminant Analysis classifier which achieved an accuracy of 72 ± 3 % on 19 human subjects. Although the studies provide interesting and relevant results, they are all limited to being a newly developed separate device that would have to be integrated into an already demanding and time-consuming surgical procedure. In contrast, tissue classification using the ultrasonic aspirator does not encounter the same limitations, as the tissue resection process is not hindered and classification is based on the data obtained. This would allow for tissue differentiation to be integrated into an existing workflow and not lead to an impairment.

Outside the realm of brain tissue, different approaches are investigated. The work of Gutierrez-Giles et al. [67] focuses on tissue classification without the use of a force/tactile and velocity sensor in a robotic palpation scenario. They estimated these properties by feeding the robot's joint position and joint torque into a model-based state observer. Mechanical parameters of four different silicon rubbers[1] are estimated by a least squares algorithm based on the estimated force and velocity. Subsequently, the estimated mechanical parameters are used to classify the silicon rubber with a Bayesian classifier, which achieves an accuracy of up to 100 %. Zhang et al. [186] evaluated tissue differentiation with a piezoelectric tactile sensor in dependence of arbitrary contact angles. Resonance frequency shifts are used for differentiation between five different types of tissue samples in simulation and laboratory experiments[2]. The authors optimize the sensors parameter such that the influence of the contact angle is minimized on the resonance frequency shift. Contact angles of 0°, 15°, 30° and 45° are used throughout the experiments, which are clustered with a k-means algorithm in five clusters for the five employed tissue types. Results show a correct rate of clustering of 92 %, indicating that the sensor is robust against contact angle changes.

Electrical impedance spectroscopy is used by Kent et al. [91], which analyzes the impedance of tissue samples at different frequencies. A bipolar electrode is used in a modified brachytherapy needle with an impedance spectroscopy system to measure the impedance of liver, kidney, muscle, testicle, and ovary tissue from poultry, bovine, and canine origin. Different feature selection methods are compared to a modified forward stepwise selection algorithm in combination with a variety of classifiers. The authors demonstrated that their feature selection method performance was on par with the unmodified forward stepwise selection algorithm with an accuracy of 86 ± 7 %. However, they were able to further reduce the number of features with their method,

1 EcoFlex and Dragon Skin by Smooth-On, Inc. Young's moduls at least 50 000 Pa [128].

2 Young's modulus of at least 45 000 Pa in simulation and Shore hardness of at least 00-24 in laboratory experiments. Brain Shore hardness: 00-4.5 [172].

which can be beneficial for the real-time application of the system. Similarly, Youn et al. [185] used different types of subdermal electrodes to measure the tissue impedance of six different types of tissue (muscle and fat from neck and abdomen, liver and lung) from porcine origin. A multi-classifier system is used to classify the tissue types based on the impedance measurements at different frequencies. Each classifier's features are selected by a genetic algorithm approach in order to optimize the classification performance. A final neural network combines the results of the individual classifiers to achieve an accuracy of 98.6 %, significantly outperforming the individual classifiers. Furthermore, robustness was improved by introducing white Gaussian noise to the impedance measurements in follow-up work [184].

Differentiating between tissue types using a vibro-acoustic sensing approach is investigated by Sühn et al. [167]. To this end, a standard palpation probe is equipped with a vibration measurement system to capture the vibro-acoustic signals generated by palpating the instrument manually at 45° over a distance of 3 cm on the tissue's surface. Six different tissue types are employed in the experiments, including three non-animal tissues (wood, felt, gel) and three animal tissues (bone, tendon, liver). Signals are processed via a Continuous Wavelet Transform (CWT) for a time-frequency spectrum, followed by instantaneous dominant frequency extraction, its probability distribution and a CWT-based averaged spectrum. Subsequently, feature extraction is performed and a classification is conducted with a support vector machine and a k-nearest neighbor classifier. The results reveal an accuracy of at least 93.8 % for the tasks of classifying three tissues and a reduced accuracy of 78.8 % for the classification of all six tissue types. Instead of leveraging vibro-acoustic signals, the work of Ostler et al. [126] focuses on intra-abdominal acoustic signals from a microphone inserted via trocars in a laparoscopic setting of a box trainer. Electrosurgery was performed in the box trainer on porcine liver, muscle, fat, and fascia tissue. The acoustic signals were processed with a sliding window technique to compute log-mel-spectrograms, serving as input for a convolutional neural network. Differentiation results reached an accuracy of 89.9 % for tissue classification and 86.4 % for additional electrosurgical mode differentiation.

Within the realm of skin cancer, Santilli et al. [152] investigate the feasibility of using an intelligent knife, iKnife, for real-time classification of basal cell carcinoma margins. The iKnife is a surgical tool that combines electrosurgery with rapid evaporative ionization mass spectrometry to analyze the smoke generated during electrosurgical dissection. 190 ex vivo samples (63 basal cell carcinoma and 127 healthy skin tissues) were acquired and processed with a deep learning-based combination of an autoencoder and classification network and a data augmentation approach. Classification results showed an accuracy of 96.6 ± 1.3 % for the differentiation of basal cell carcinoma and healthy skin tissue. This work is further extended by Jamzad et al. [80], which incorporates an attention-based multiple instance learning model to the classification task and evaluates the influence of different tips of the iKnife. Similar performance to the previous work was achieved, while dependency of tip type on the classification performance was

observed. Additionally, Janssen et al. [82] investigated the clinical feasibility of using the iKnife for real-time navigation during basal cell carcinoma surgery, showcasing the potential in intraoperative decision-making evaluations. Furthermore, domain adaptation from skin cancer to breast cancer and self-supervised learning approaches are conducted by Santilli et al. [153, 154] with the iKnife, allowing for the classification of breast cancer margins with an accuracy of up to 97 %. Within the same scope of skin cancer, Bratchenko et al. [16] conducted a study on the in vivo classification of skin cancer using a combination of Raman and autofluorescence spectroscopy in a portable low-cost device. Classification is carried out using Partial Least Squares Discriminant Analysis by constructing latent variables to discriminate between skin cancer entities based on spectral differences in the Raman and autofluorescence spectra. The results are presented as receiver operating characteristic curves, which demonstrate an area under the curve of 0.75 for differentiation, indicating that the model performs comparably to or even better than humans. Unlike previous studies, the authors of this research did not prioritize margin detection; instead, they concentrated on classifying the type of skin cancer itself. Other optical methods, using polarimetry, are investigated by Van Eeckhout et al. [48] for tissue classification. The authors evaluated five depolarization metrics on 120 samples from poultry tendons, muscles, and myotendinous junction regions. Classification was carried out with a decision-tree classifier, linear discriminant analysis and a k-nearest neighbor classifier to identify the most suitable depolarization metric spaces for tissue differentiation at different wavelengths. The k-nearest neighbor classifier achieved the best performance with accuracies of up to 65 %. In extension to this work, Canabal-Carbia et al. [27] investigated the suitability and potential superiority of indices of polarimetric purity as depolarizing observables for biomedical applications, demonstrating their ability to enhance contrast between tissue structures and discriminatory power.

Generally, all these approaches are primarily geared toward robotic applications and minimally invasive surgery. Furthermore, they involve newly developed devices that necessitate workflow adjustments and additional training for surgeons. Also, consideration needs to be given to the fact that biological tissue and articifial tissue used in the studies have different mechanical properties in comparison to brain tissue, which is several orders of magnitude smaller. Thus, it is important to note that direct translation to neurosurgical practice is limited. Regarding ultrasonic aspirators, different studies have shown that a commercially available ultrasonic aspirator – in addition to its primary function of tissue removal – can be used in different ways without any changes to the surgical workflow. The aspirated tissue for example, which is usually considered as biological waste, can be used for culturing glioblastoma stem cells [8], thus suggesting the ultrasonic aspirator as a tissue sampling tool. This tissue sampling capability was investigated with a sequential mass spectrometry approach by Agar et al. [4]. A desorption electrospray ionization mass spectrometer was used to analyze the in vivo

aspirated tissue samples after five minutes of drying under nitrogen and transferring to a glass slide. Only qualitative results were obtained, showing the potential of the ultrasonic aspirator as a tissue sampling tool for mass spectrometry. Advances in this field were made by Schäfer et al. [155], who used a similar approach to Agar et al. [4] but with a different mass spectrometer. They combined the ultrasonic aspirator with a Venturi easy ambient sonic-spray ionization mass spectrometer to analyze the aspirated tissue samples in real-time. Analysis was carried out on the spectra of the samples using a principal component analysis and a linear discriminant analysis. 284 ex vivo human brain tissue spectra were analyzed, showing a 100 % distinction between glioablastoma and healthy brain tissue. However, a detailed quantitative analysis was not provided. An overview of the related work on tissue differentiation using various methods and materials with their respective performance is given in Table 6.1.

In contrast to the aforementioned approach of processing biological waste, the subsequent sections of this chapter examine the use of the ultrasonic aspirator at an earlier stage, by utilising the electrical data obtained during the interaction between the instrument and the tissue.

Table 6.1: Overview of related work on tissue differentiation using various methods and materials with their respective performance.

Paper	Method	Materials	Performance
[170]	Fluid-filled balloon sensor system	Urethane gel, white and gray matter of porcine brain	Qualitative assessment
[85]	Piezoelectric tactile sensor with torsional resonators	In vivo rat brain tissue	Correlation coefficients up to 0.87
[165]	Piezoelectric bimorph sensors	Ex vivo porcine brain tissue	Qualitative assessment
[112]	SEEG data	In vivo gray and white matter of human brain	72 ± 3 % accuracy
[67]	Model-based state observer for robotic palpation	Four silicon rubbers	Up to 100 % accuracy
[186]	Piezoelectric tactile sensor with resonance frequency shift	Five tissue samples in simulation and lab	92 % correct clustering rate
[91]	Electrical impedance spectroscopy	Liver, kidney, muscle, testicle, ovary tissue from various animals	86 ± 7 % accuracy
[185]	Impedance measurements with subdermal electrodes	Muscle, fat, liver, lung from porcine	98.6 % accuracy
[167]	Vibro-acoustic sensing with palpation probe	Various non-animal and animal tissues	At least 93.8 % accuracy
[126]	Acoustic signals from electrosurgery in laparoscopic setting	Porcine liver, muscle, fat, fascia tissue	89.9 % accuracy
[152]	iKnife with ionization mass spectrometry	Ex vivo basal cell carcinoma and healthy skin	96.6 ± 1.3 %
[16]	Raman and autofluorescence spectroscopy	In vivo skin cancer	Area under the curve of 0.75
[48]	Polarimetry with depolarization metrics	Poultry tendons, muscles, myotendinous junctions	Up to 65 % accuracy
[4]	Mass spectrometry with ultrasonic aspirator	In vivo human brain tissue	Qualitative assessment
[155]	Mass spectrometry with ultrasonic aspirator	Ex vivo healthy human brain tissue and glioblastoma	100 % distinction

6.2 Feasibility under Controlled Conditions

6.2.1 CNC Controlled Differentiation Without Tissue Ablation[1]

In a first step, the feasibility of using electrical data that is recorded using controlled conditions with an ultrasonic aspirator to differentiate between different tissue types is investigated.

Data Preparation For this purpose, the data set D1 is used. This data set is recorded with an ultrasonic aspirator on a CNC machine with settings set to a non-resecting mode, i.e. the ultrasonic aspirator does not ablate any tissue with aspiration and irrigation turned off and ultrasound amplitude set to the minimum. Additionally, a sonotrode is used without an aspiration channel, which is not the case in a clinical setting where tissue aspiration is of key importance.

Methods Within this study, the goal is to solve a regression task, where the PVA ratio component of the artificial tissue models is predicted. The PVA ratio is a surrogate measure for the stiffness of the tissue and thus applicable for the differentiation of tissue types within the scope of this study. Electrical features are used as input for the regression task. In order to perform a regression on this data, three empirically determined nonlinear methods are used. The first two methods are deep learning-based regression models: a Fully-Connected Network (FCN) that takes the features as input to regress to the respective stiffness value and a 1D residual network (ResNet) which uses a sliding windows approach with a window size of four seconds to regress to the center point of the window [72]. The initial approach, which is based on the analysis of the current data, does not consider temporal information from adjacent data points. In contrast, the latter method aims to enhance performance by leveraging temporal information from the surrounding data. The last method uses a Gaussian Process (GP) with an exponential kernel to determine the stiffness value. An advantage of using GPs for a regression is the incorporation of uncertainty into the predictions, which can be beneficial in later applications. A five-fold cross-validation is conducted over the 30 recordings with a total of more than 11 000 data points. The same splits are used across the different methods and normalized to have zero mean and unit variance. Data points without any contact to the tissue models are assumed to have a stiffness value of zero. To evaluate the performance of the methods, the metrics root-mean-square error (RMSE), mean absolute error (MAE) and R^2 are obtained and averaged over the five folds.

Regression Performance The summarized results of the regression can be found in Table 6.2. For all methods, a high R^2 value of more than 0.9 can be obtained. Since the

1 Parts of this section are based on the publication [11].

two deep learning-based methods show similar results, it can be deduced that there is limited impact of temporal information on the predictions. The GP regressor model yields superior performance compared to the deep learning-methods. All methods exhibit high standard deviations in the RMSE and MAE metrics, indicating considerable fluctuations in predictions around the true stiffness values. A qualitative result of the GP can be found in Figure 6.2. Particularly noteworthy are the large uncertainties in the area of the jump discontinuities of the ground truth signal and the samples around. Those jump discontinuities are state changes which are caused by bringing the instrument into contact with the tissue model and removing it again. In Figure 6.3, an exemplary voltage signal of a recording without tissue resection is shown on the left. It can clearly be seen that the signal is overshooting at the beginning and end of the *contact* phase, due to the system's response to the sudden change in load. This overshooting of signals is a challenge for the regression models and can impede the learning process. However, the communication of the uncertainty of the model's prediction to the user allows rejecting the prediction in a later clinical application, which can be particularly relevant in a safety–critical environment such as the neurosurgical operating room.

Table 6.2: Performance comparison of regression methods without tissue ablation. The best value of each metric is highlighted.

Method	RMSE	MAE	R^2
FCN	0.894 ± 0.619	0.473 ± 0.409	0.906 ± 0.106
ResNet	0.823 ± 0.870	0.454 ± 0.621	0.903 ± 0.188
GP	$\mathbf{0.561} \pm 0.539$	$\mathbf{0.286} \pm 0.365$	$\mathbf{0.940} \pm 0.092$

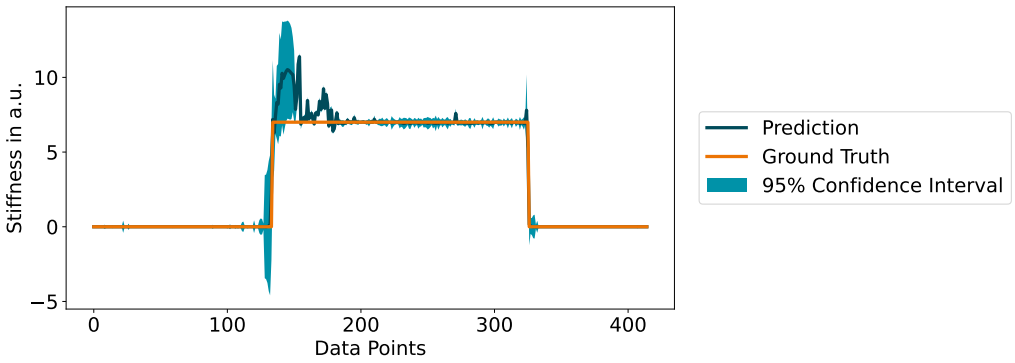


Figure 6.2: Qualitative result of the GP regression. The blue line represents the true stiffness value, the orange line the predicted value and the shaded area the uncertainty of the prediction.

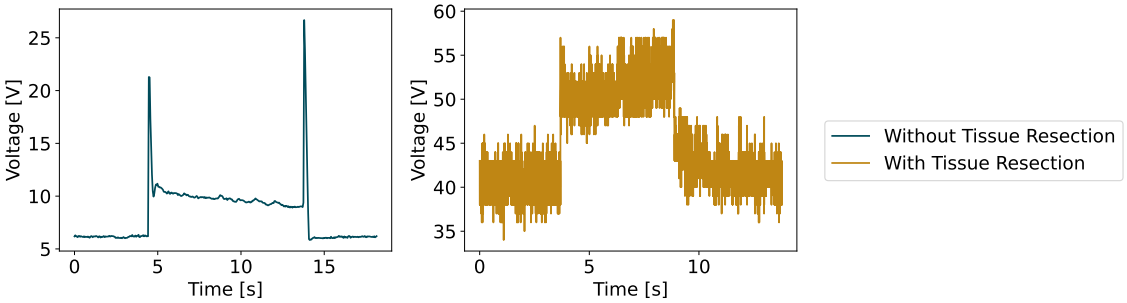


Figure 6.3: Exemplary voltage signal of a recording without and with tissue resection on similar tissue models. The signal without tissue resection (settings: no aspiration and irrigation, ultrasound set to minimum) enabled on the left shows less noise but distinct overshooting of the signal at the beginning and end of *contact* phase. The signal with tissue resection (standard settings) on the right shows more noise but reduced overshooting of the signal.

6.2.2 Differentiation under Tissue Ablation in a CNC Setting[1]

In a second step, the feasibility of tissue differentiation with an ultrasonic aspirator is investigated under the means of tissue ablation.

Data Preparation The data used for this investigation is the data set D2, which is recorded with a CNC machine to keep the experimental conditions as controlled as possible. All data are sampled with a sampling frequency of 1000 Hz. A single, appropriate setting is used for the generator to ensure a sufficient ablation of the tissue models. For this study, tissue models are used with two different mechanical properties. The Young's modulus values 1513 Pa and 5709 Pa are used as categorical targets for the classification in the subsequent experiments. Thus, focus is shifted from a regression task to a classification task to reduce the complexity of the task, address the high standard deviations in RMSE and MAE metrics that indicate considerable fluctuations in predictions, and minimize the influence of the inherent uncertainty of the stiffness values of the tissue models caused by the manufacturing process. Additionally, the inherent uncertainty of stiffness will be more problematic the more diverse the data set gets, which is easier to cope with in a classification task.

As illustrated on the right side in Figure 6.3, a slight increase in the signal over time during contact can be observed. This is likely due to the generator's slow adaptation to the sudden load change. Additionally, direct feedback is desired for later clinical use, and therefore the system is trained so that only a small amount of time is required to allow for classification. Thus, only the first second of the *contact* phase is used, leading to 1000

1 Parts of this section are based on the publication [13].

measurement points per recording. This results in a total of 40 000 measurement points that are processed and classified individually (hereinafter referred to as sample-based classification) with respect to their tissue stiffness classes 1513 Pa and 5709 Pa. Input features to the classification models are the electrical parameters and each of its relative shifts. For this, the mean value of the first 30 points before contact is subtracted from the measured points and divided by the same value to obtain a relative shift as features per electrical parameter. These additional hand-crafted features can be seen as an individual scaling of the data points to the initial state of the *contact* phase.

During data analysis, a significant portion of recordings show a superimposed noise signal with a frequency of around 1.2 Hz, possibly caused by periodic behavior of the irrigation function of the instrument, which could impede classification performance. This noise is related to fluid buildups in the cover of the sonotrode, which is used to cool the instrument during tissue ablation while also maintaining tissue hydration. Two different preprocessing approaches are applied to filter this noise: a low-pass filter using a Butterworth filter, leading to a smooth signal, and band-stop filter that uses a notch filter allowing the data to keep high-frequency components. The *contact* phase of one recording is shown in Figure 6.4, which illustrates the influence of the two preprocessing approaches on the raw signal with periodic noise. The design choices of the two filters were determined empirically to reduce the occurrence of noise at 1.2 Hz. The low-pass Butterworth filter was set to a third-order filter using a cut-off frequency of 0.75 Hz and a sampling frequency of 1000 Hz. The band-stop notch filter's setting showed an overall sufficient result using a quality factor of 0.4 with the frequency to remove set to 1.2 Hz at a sampling frequency of 1000 Hz. A quality factor measures the selectivity of the filter, with a higher value leading to a narrower frequency band to be attenuated. It needs to be noted that those filters are applied on the entire signal after recording and not in an online fashion. The use of filters in an online fashion will always entail a time delay which is inevitable and must be taken into account accordingly.

Methods Classification of the different tissue models based on the electrical features of the ultrasonic aspirator is performed with three different classification algorithms. First, a random forest (tree size: 100, feature importance: Gini index) is used as a baseline model [17]. This algorithm is chosen because it provides a robust baseline with its simple and comprehensible design. The second algorithm is utilizing a neural network (NN) structure with a combination of an autoencoder and a classifier network. The autoencoder is used for regularization purposes to enforce the NN to only learn useful properties of the data in the latent space. Finally, to classify the data, a classification network is used, whose input is based on the latent space. Both, the autoencoder and classification network are trained simultaneously end to end with a combined loss

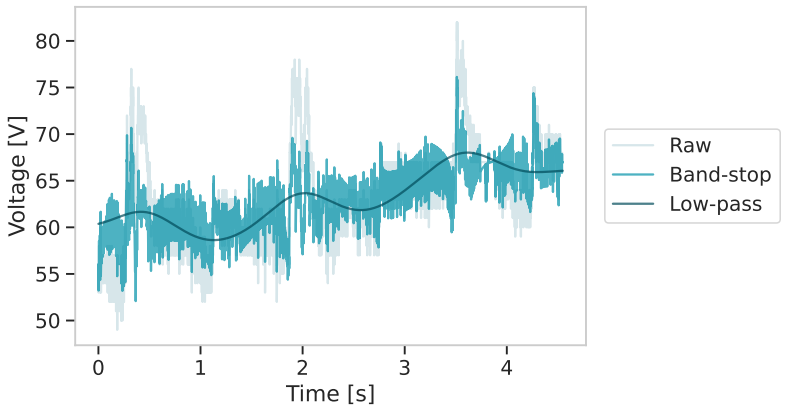


Figure 6.4: Influence of the two preprocessing filters on the raw signal with periodic noise. The raw signal is shown in light blue, the low-pass filtered signal in dark blue and the band-stop filtered signal in medium blue.

inspired by Santilli et al. [152]. The combined loss is defined as:

$$loss = loss_{clf} + \alpha \cdot loss_{ae} \tag{6.1}$$

with $loss_{clf}$ as the cross-entropy loss of the classification network and $loss_{ae}$ as the mean-squared error loss of the autoencoder. The effect of the $loss_{ae}$ on the combined loss is controlled by the weight parameter α. Throughout the experiments, α is set to 0.3, which was empirically determined to be a good general fit. The autoencoder consists of one fully connected hidden layer of size 32 in the encoder and decoder parts with one latent space layer of size nine in between. The classification network features one hidden layer of size four as well as one output layer. Table C.1 provides a more detailed overview of the parameters used, while the network architecture is shown on the left in Figure 6.5. Both classifiers, the RF and NN, work on a sample-based level; thus, no temporal characteristics of the signals are leveraged. In order to incorporate temporal information into the training process, a deep convolutional neural network (CNN) with a sliding-window approach is used as a third classification algorithm. The CNN uses a 1D residual network architecture (ResNet) that allows an ease of training due to residual blocks with identity skip-connections [72, 71]. A total of seven convolutional layers are used, formed by three residual blocks. With the beginning of every residual block, the numbers of feature channels are doubled, while a stride size of two is applied – with an exception for the second residual block in which the feature channel number is not doubled. The kernel size is fixed to seven on all layers and the number of the first feature channels is set to 16. Every convolutional layer is followed by a 1D batch normalization layer. After the last convolutional layer, an average global pooling is performed to enable a transition to the fully connected output layer. The input length to the CNN

is based on the window that is moved over the signal with a size of 64. Prediction is always performed on the last value in the window and only done if 63 previous values are available. Cross-entropy loss is chosen as the loss function for this architecture. An overview of the architectural design is provided on the right in Figure 6.5 with its parameters given in Table C.2.

Training for both neural network approaches is done with Adam optimization [95] using an initial learning rate of 0.01, which is multiplied by a factor of 0.1 upon reaching a plateau on the training data. Early stopping is implemented based on validation data performance with an otherwise maximum epoch count of 2000. The batch size is set to 16 384, and a weight decay of 0.001 is applied. Input data are normalized with zero mean and unit variance. In order to manage overfitting, dropout is used in case of NN in the hidden layers of the encoder and classification networks, as well as after every convolutional layer in the CNN. In both cases, the dropout value is set to 0.3. Furthermore, for the NN a Gaussian noise of 0.1 is added to the input features of the training data to prevent overfitting. During inference on the NN, the decoder part of the autoencoder is omitted and only the classification output is used.

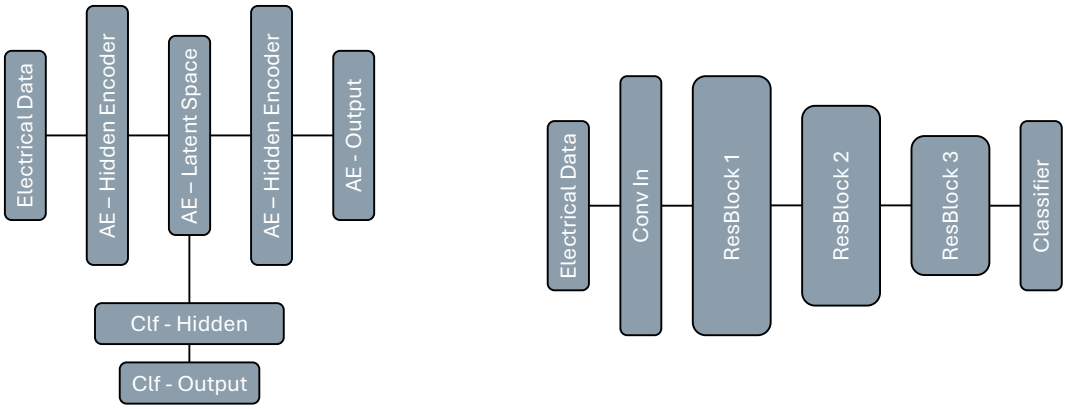


Figure 6.5: Network architecture of the neural network (left) and the convolutional neural network (right). Details of the hyperparameters can be found in Table C.1 and Table C.2, respectively. AE: autoencoder, Conv: convolution, Clf: classifier, ResBlock: residual block.

Experiments are carried out in two steps: First, the performance of the aforementioned machine learning approaches is evaluated; second, the influence of data preprocessing on classification is analyzed. All experiments are conducted and analyzed over a five-fold cross-validation. It is ensured that the splits do not divide the recording data, such that data from the same recording are not found in training and test sets. For the neural network-based classifiers, an additional split of 87.5 % and 12.5 % is performed for training and validation data, respectively, to prevent overfitting. Different metrics are

recorded on the test data for evaluation purposes: the F1-score (F1), accuracy (ACC), positive predictive value (PPV), true-positive rate (TPR) and the area under the receiver operating characteristic curve (AUROC). As a first experiment, the performance of the three presented classification methods, RF, NN, and CNN is evaluated without any prior signal processing. All methods are provided with the same electrical input signals, from which a differentiation of the tissue models between the stiffness values is then to be carried out. The second experiment concerns possible data preprocessing methods to filter out the periodic noise at 1.2 Hz. The influence of the two preprocessing filters is evaluated with regard to the classification performance of RF, NN, and CNN.

Classification Performance Table 6.3 presents the results of the classification performance of the three proposed methods. The sample-based approaches RF and NN both show equivalent performance with a mean F1 of about 0.68. A slightly higher mean F1-score of 0.72 is obtained with the CNN approach. However, reviewing the standard deviations for all five metrics indicates that differences in metrics might be negligible. Overfitting is observed during training and is most likely caused by the high variance of the data. This is why multiple countermeasures are implemented to keep the effects at a minimum.

Table 6.3: Performance metrics of classifiers with tissue ablation in a CNC setting. The best value of each metric is highlighted.

Classifier	F1	ACC	PPV	TPR	AUROC
RF	0.688 ± 0.206	0.692 ± 0.203	0.700 ± 0.212	0.692 ± 0.203	$\mathbf{0.790} \pm 0.220$
NN	0.682 ± 0.181	0.687 ± 0.175	0.689 ± 0.182	0.687 ± 0.175	0.735 ± 0.211
CNN	$\mathbf{0.720} \pm 0.122$	$\mathbf{0.735} \pm 0.104$	$\mathbf{0.773} \pm 0.097$	$\mathbf{0.735} \pm 0.104$	0.778 ± 0.224

Influence of Data Preprocessing Two preprocessing filters (Figure 6.4) are applied to handle occurring periodic noise by smoothing the signal with a low-pass filter or by explicitly filtering the corresponding frequency using a band-stop filter. The results are provided in Table 6.4. In general, the use of both preprocessing approaches leads to an improvement of the classification results. In the best case, the performance of the NN is increased from 0.682 ± 0.181 mean F1 on raw data to a mean F1 of 0.900 ± 0.096 using a low-pass filter. The noise apparent in the raw data thus seems to have an influence on the quality of the classification, which could be attributed to the wide range of signal values, as exemplified in the noise in Figure 6.4. With regard to the two filters, it can clearly be observed that low-pass preprocessing leads to superior results in direct comparison with band-stop preprocessing. This could be due to the reduced complexity of the data space by smoothing out signals through which the learning of the classification models is simplified and more robust. Concerning only the low-pass

filter, the metric is boosted by at least 10 percentage points across all methods with respect to the results based on raw data input. In the case of band-stop filtering, a maximum increase of 5 percentage points can be observed, with a slight deterioration of the metrics for RF. With respect to the CNN results that include temporal information into the predictions, it can be seen that a slightly increased performance can be achieved with the low-pass filtered data, although it does not contain high frequencies compared to the band-stop-filtered data. Even with data being sequential in its recording, the results indicate that the temporal information is not a relevant factor for the classification task at hand and can be improved with adjusted data filtering using a sample-based method such as the NN by a large margin. Using preprocessing, RF performance is inferior to both neural network-based approaches. Since filtering has simplified the data space, it is arguably easier for the neural network-based approaches to train and perform an improved feature extraction. This feature extraction may be the reason for the difference between the approaches, as the RF classifier is limited to the fixed hand-selected features.

Table 6.4: Performance metrics of classifiers with tissue ablation in a CNC setting with different preprocessing filters. The best value of each metric is highlighted.

Classifier	Filter	F1	ACC	PPV	TPR	AUROC
RF	Low-pass	0.788 ± 0.238	0.795 ± 0.229	0.805 ± 0.232	0.795 ± 0.229	0.881 ± 0.170
RF	Band-stop	0.658 ± 0.201	0.664 ± 0.199	0.681 ± 0.214	0.664 ± 0.199	0.748 ± 0.245
NN	Low-pass	$\mathbf{0.900} \pm 0.096$	$\mathbf{0.902} \pm 0.092$	$\mathbf{0.918} \pm 0.070$	$\mathbf{0.902} \pm 0.092$	$\mathbf{0.935} \pm 0.090$
NN	Band-stop	0.703 ± 0.199	0.710 ± 0.190	0.714 ± 0.197	0.710 ± 0.190	0.754 ± 0.232
CNN	Low-pass	0.828 ± 0.192	0.838 ± 0.173	0.849 ± 0.159	0.838 ± 0.173	0.838 ± 0.291
CNN	Band-stop	0.799 ± 0.146	0.801 ± 0.144	0.816 ± 0.147	0.801 ± 0.144	0.844 ± 0.227

6.2.3 Contextualization and Synthesis of Cumulative Findings

Tissue differentiation with an ultrasonic aspirator was evaluated in a two-step process under controlled conditions using a CNC machine. The first step investigated the general feasibility of tissue differentiation with electrical data from an ultrasonic aspirator without the additional difficulty of tissue ablation. This study showed the possibility to learn a relationship between the electrical data and the stiffness of the tissue models. Values of more than 0.9 for the R^2 score indicate good performance of all models. However, temporal influence were only limited on the performance of the deep learning-based methods, which is not surprising, as the data did not change drastically over time in its respective *contact* and *no contact* phases, giving no benefit to temporal information. As apparent from the results, overshoots in the signal posed a challenge for all the models. Due to the system's response, those overshoots were likely only observed in the data because of the experimental setup with modified settings and diminished strongly in the data with tissue resection, as shown in Figure 6.3. Regression as a starting point of the investigation proved as an adequate choice, for simplicity and to get a first impression

of the feasibility of the task. Predicting the chemical ratio of the tissue model within the scope of this study was sufficient on artificial tissue models, as it is a surrogate measure for the stiffness of the tissue. However, to eliminate possible sources of error, such as changes to the manufacturing process of the tissue models, the focus needed to be shifted to the actual stiffness in terms of Young's modulus, which otherwise would get more complex with a higher diversity of the data set. Within the context of this study, the feasibility of tissue differentiation with an ultrasonic aspirator was shown under controlled conditions. Obviously, these controlled conditions are not representative of a real surgical scenario, as the ultrasonic aspirator is not used in a handheld mode and no tissue ablation is performed. Additionally, the used sonotrode is not equipped with an aspiration channel, which is a necessity in a clinical setting. Thus, the next logical step was to investigate the feasibility of tissue differentiation under tissue ablation to increase the difficulty of the task and mimic a real surgical scenario step by step.

This logical next step was taken in the second part, where the feasibility of tissue differentiation with an ultrasonic aspirator under tissue ablation was investigated – still in a CNC setting. The results of this study showed, that even under the more challenging conditions of tissue resection, the differentiation of tissue models with an ultrasonic aspirator was feasible. An important part of the results and experiments was the filtering of the noise in the signals. This periodic noise was most likely caused by the irrigation of the instrument. It is apparent that the noise has a strong impact on the classification performance, as the results showed a clear improvement in the classification performance after filtering the noise. Thus, in addition to the filtering of the noise, hardware related measures to reduce the noise in the signals are necessary. Keeping in mind that the filtering was done on the entire signal after acquisition, it is not applicable in a real-time setting. Again, a temporal influence seemed to have no benefit for a classification task, as the sample-based approaches showed similar performance to the CNN. This might be due to the similar behavior during the *contact* phase of the signal. This is also supported by the fact that the low-pass filtered signal without high-frequency components led to a better performance than the band-stop filtered signal. Thus, no additional information was gained through the temporal information of the signal, which should have been pronounced in higher frequency components.

As the data that were processed are only the beginning of the *contact* phase, prior information about the state of contact was necessary. Since this data set was postprocessed to label *contact* and *no contact* phases, this information was already available and easy to use. Within this study this was a legitimate approach, as the focus was on the feasibility of the task and not its real-time applicability. However, in a real-time setting, this information needs to be obtained from the electrical data themselves and processed in real-time as shown in the previous chapter.

This study did not consider the influence of the generator setting on the classification performance, nor did it consider the influence of a handheld mode. Generator settings will have direct influence on the electrical data even without any contact to the tissue.

For example, a higher ultrasound setting will directly influence the voltage signal since a higher amplitude is used. This can be seen in Figure 6.6 for an instrument without contact to the tissue in different generator settings.

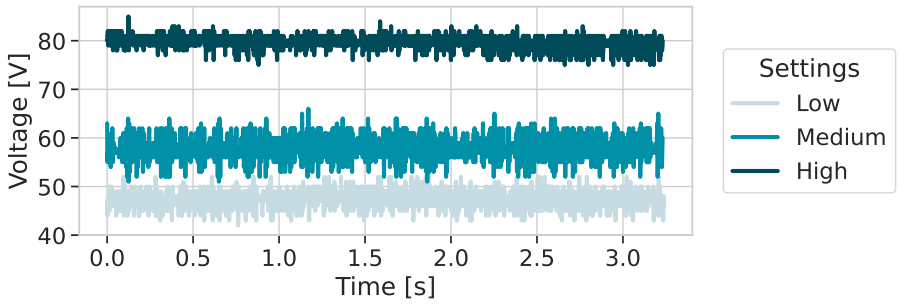


Figure 6.6: Voltage signal of an ultrasonic aspirator without contact to the tissue in different generator settings. The higher the generator setting, the higher the amplitude of the signal.

Similarly – as shown with the irrigation noise – fluid flow through the instrument will have an influence on the electrical data, that are also influenced by the generator setting. More fluid flow will lead to a higher load on the instrument and therefore a higher power demand on the generator. Those steps are covered in the following sections.

6.3 Feasibility under Realistic Conditions

Seeing that tissue differentiation is feasible under controlled conditions, the next step is to investigate the usage under increasingly realistic conditions. For this, two data sets with recordings from a handheld mode are used with different generator settings. In a first step, the influence of generator settings on the prediction is investigated, while at the same time initial results for handheld use are obtained. This step is executed with the data set D5, which consists of a limited number of recordings with different generator settings for preliminary assessment. The second step increases the data variety and possible tissue classes as well as features used for classification. With regard to the observed influence of the irrigation noise, the cover of the sonotrode is modified such that a fluid buildup is not possible. Therefore, the periodic noise is not present in the data set and no preprocessing is necessary. For this step, the data set D6 is used, which consists of a larger number and more diverse recordings in terms of tissue models and generator settings. In a third step, classification is moved from a flat classification to a hierarchical classification, as it is observed that some targets are frequently misclassified.

6.3.1 Influence of Generator Setting[1]

Data Preparation Three synthetic tissue models, different in their stiffness, are used (D5). Data from each tissue model is recorded with two different generator settings, referred to as *low* and *high* and preprocessed with the same low-pass filter as in the previous section. A total of 97 contacts is recorded for which the first half-second of *contact* is used, resulting in 33 950 measurement points. This short time limitation reduces potential influences of the handheld movement and other fragmentation effects to a minimum. All contact initializations of the recordings are isolated and used for feature generation. As before, hand-crafted features are used for classification, which are based on the relative shift of the electrical parameters for each single time step. Training and test splits are defined, resulting in 87.5 % and 12.5 % of the data respectively. Care is taken to not split measurements from the same contact initialization into training and test sets.

Methods Classification is done with two machine learning approaches: an AdaBoost and a neural network (NN) classifier. The AdaBoost classifier is chosen as a baseline model with a maximum depth of one, a learning rate of 0.1 and a total of 125 estimators. The NN classifier architecture is the same as used in the previous section and consists of an autoencoder and a classification network part. This combination allows for regularization of the NN and the enforcement of learning useful properties of the data in the latent space. Utilizing the same combined loss as depicted in Equation 6.1, α is set to one. The autoencoder's hidden layer is adjusted to a size of 16, while the latent space is defined with six dimensions. As for the classification part, the hidden layer consists of a size of four, with three for the output layer. The same learning routine as in subsection 6.2.2 is used for training the NN with an Adam optimizer [95], an initial learning rate of 0.01, a batch size of 2048, epoch length of 500 and a weight decay of 0.001. Since the settings of the generator are set manually on the generator, they are known and used as prior information for classification, meaning that for each generator setting a separate classifier is trained on their respective data. This results in two setting-based classifiers, in addition to one general single classification model over all generator settings. All methods process data on the sample level, meaning that no temporal characteristics of the signals are considered due to limited influence on previous results. Results are given in the metrics F1-score (F1), accuracy (ACC), positive predictive value (PPV) and true-positive rate (TPR), and are evaluated on the test data.

Generator Setting Dependent Classification The results for both variants of the classification approach – the general single classifier and the setting-based classifiers –

1 Parts of this section are based on the publication [12].

are shown in Table 6.5. They clearly indicate a difference in the performance of the

Table 6.5: Classifier performance for tissue differentiation under different generator settings. The best value of each metric is highlighted.

Classifier	ACC	F1	PPV	TPR	Settings
AdaBoost	0.755	0.748	0.774	0.757	low & high
NN	0.910	0.910	0.914	0.910	low & high
AdaBoost	0.814	0.799	0.880	0.814	low
AdaBoost	0.887	0.885	0.895	0.887	high
NN	**0.943**	**0.944**	**0.951**	**0.943**	low
NN	0.935	0.934	0.938	0.935	high

classification approach, showing that the usage of prior information of the generator settings can improve the results regardless of the classification method. This is particularly evident when looking at the distribution of an exemplary feature, such as voltage, in Figure 6.7. In case of a general single classifier which does not take the generator setting

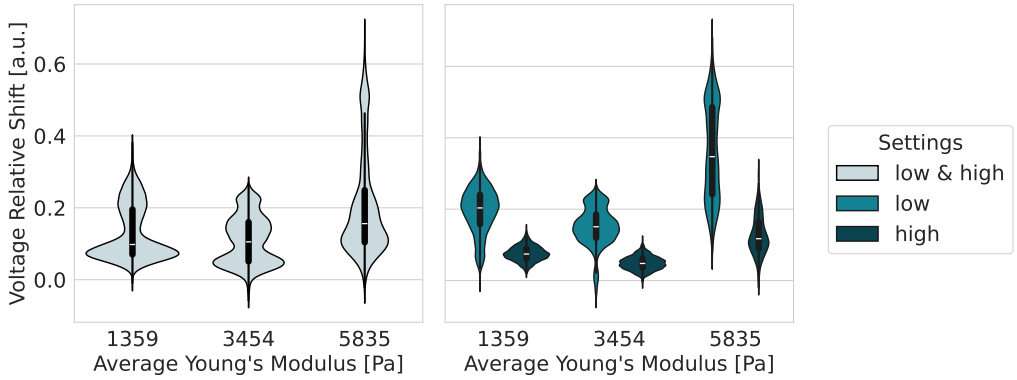


Figure 6.7: Exemplary distribution of the voltage feature for the different generator settings. On the left side, the distribution is shown if no distinction is made between the generator settings, on the right side the distribution is shown if the generator settings are taken into account. A setting dependent distribution reduces overlap and thus improves classification potential.

into account, the distribution of the voltage feature over the Young's modulus tends to overlap. However, when the generator settings are taken into account, the distribution of the voltage feature is more separated, which is beneficial for the classification task. For example, the voltage feature of Young's modulus 3545 Pa for *low* shares a similar appearance to 5835 Pa for *high*. With the inclusion of the setting information in the training, this overlap is reduced and the classification performance is improved. It has to be noted that classification is not only dependent on the single voltage feature, but on a combination of features and is only used for illustrative purposes. With regard

to the overall classification performance concerning the classification models, the NN outperforms the AdaBoost classifier by more than 10 percentage points in ACC and F1. For the setting-based results, a performance difference is observed between the low and high generator settings only for the AdaBoost classifier. In case of the NN classifier, an on par performance is observed for both settings. The results indicate the feasibility to differentiate tissue models while fragmenting them with an ultrasonic aspirator under realistic conditions in a handheld mode. Nevertheless, this study is limited to a small number of tissue models and individual contacts. Additionally, only simple hand-crafted features were used. A more diverse data set is necessary to capture the variety present in a real intraoperative setting, as well as a more sophisticated feature extraction method to capture the complexity of the electrical data.

6.3.2 Feature Extraction and Data Size Increase

Data Preparation In the previous subsection 6.3.1, the influence of the generator setting on the classification performance is investigated. The results show the importance of the generator setting for the classification performance, as well as the feasibility of tissue differentiation under realistic conditions. However, the data set used is limited in its diversity and size, and feature extraction is based on simple hand-crafted features for each temporal step of the *contact* phase of the signal. To address this issue, a new feature extraction method is employed, and the data set size is expanded to better reflect the variety found in real intraoperative settings. The data set used for this step is the data set D6, which consists of a larger and more diverse set of recordings, utilizing a total of three settings *low*, *medium* and *high*. Furthermore, this data set is recorded with a modified cover for the sonotrode, which prevents fluid-dependent noise in the electrical data and thus no longer requires filtering of the noise. Due to the greater variety of data, the data are binned according to the optimal binning strategy of Bin_{opt}. This approach is well suited to the fact that there is no single stiffness value that can be assigned to brain tissue, making broader categories of ranges more appropriate [25]. Consequently, a five-class classification problem is defined and evaluated using five-fold cross-validation. In order to reduce the bias of class imbalance, the data per setting are subsampled to the class with the least amount of data. The amount is given as individual number of contacts, which for the *low* setting is 900 contacts, for the *medium* setting 850 contacts and for the *high* setting 960 contacts.

Methods Instead of a hand-crafted feature extraction method, a more sophisticated feature extraction method is used that is based on time-series data. Feature extraction is performed on the data to obtain a feature vector for each contact. This is in contrast to the previous sections in which each temporal step is predicted individually. Now, only the contact itself is classified – i.e. only the initial contact is processed. To this end, data around the contact point of the time-series data are used, which is then

standardized to achieve a zero mean and a standard deviation of one. Subsequently, the *tsfresh* library [34] is employed for automated feature extraction. The *tsfresh* library – which is short for *Time Series Feature extraction based on scalable hypothesis tests* – is a feature extraction method which extracts a large number of relevant features from the time-series data. Such features include statistical measures, time- and frequency-domain features, information-theoretic measures, and many more. Relevant features are determined by hypothesis testing to assess the statistical significance of the features. Extracted features for each contact are used as input for an extreme gradient boosting (XGBoost) classifier [30] to predict the tissue stiffness class associated with the contact for each generator setting. The XGBoost parameters are empirically determined and are set to a learning rate of one, a maximum depth of six, a maximum number of boosting rounds of 1000 and an L2 regularization term of one. Additionally to *tsfresh*, an end-to-end feature extraction and classification approach is used, which is realized with a deep learning-based residual network architecture (ResNet) [72]. This architecture, taking the same contact point data as input, is trained to predict the tissue stiffness class per contact. To allow capturing patterns early in the high level features, a larger kernel size is used in the first layer of the network followed by smaller kernel sizes in the subsequent layers, similar to the approach in [42]. A detailed overview of the architecture is given in the appendix Table C.3. Training data per fold are further split into training and validation data with a ratio of 80 % and 20 %, respectively. This allows for automatic early stopping (patience: 200 epochs) and learning rate reduction (patience: 100 epochs, reduction factor: 0.1) based on the validation data performance and aims to prevent overfitting. Training is done with Adam optimization using an initial learning rate of 0.001 [95], a maximum epoch length of 2000, a batch size of 1024, dropout of 0.1, and a cross-entropy loss. As a baseline, hand-crafted features per contact are used to compare the performance to the automated feature extraction method of *tsfresh* with the XGBoost algorithm and to the ResNet architecture. Hand-crafted features are created during the change between *no contact* to *contact* for the electrical parameters voltage and frequency as the absolute and relative shift of mean, absolute and relative shift of standard deviation, maximum derivate, full width half maximum of derivate and the slope of linear regression during *contact*. Thus, for each contact, a feature vector of 14 features is created and used as input for the XGBoost classifier. In comparison, the *tsfresh* feature extraction method extracts 1274 features per contact, while the ResNet approach extracts 64 features per contact prior to the final classification layers. Results are given in the metrics F1-score (F1), accuracy (ACC), positive predictive value (PPV), true-positive rate (TPR) and are evaluated in a five-fold cross-validation.

Feature Extraction Comparison Table 6.6 shows the results of the comparison between the three feature extraction methods per generator setting *low*, *medium* and *high*. Highest performance is observed for the *tsfresh* feature extraction method com-

pared to the hand-crafted features and end-to-end ResNet approach for all generator settings, especially for the *medium* and *high* settings. A closer look at the results in the form of a summed up confusion matrix is given in Figure 6.8 for the tsfresh feature extraction method. Even though the performance is not higher than 0.64 in mean F1, this reveals that there is a certain structure in the misclassifications. These misclassifications are mainly formed by two sections: one between the classes 500 and 1700 and one between the classes 3600, 6700 and 16 900. Misclassifications across those two sections are not zero, but are not as high as within the sections. This indicates that the classification model is unable to further differentiate within those two sections and that the classes are too similar to each other given the features and model. What is also noticeable, is that the performance seems to be correlated with the degree of the setting. A further evaluation of the results when training and testing only on a subset of binary classes given in Figure 6.9 shows that performance across the two sections is higher than within the sections. However, these intra-section binary classifications are higher than random guessing, with mean F1-scores of 0.68 to 0.84 for the *high* setting. These results therefore indicate the further away the classes are from each other, the better the classification performance is up to a 0.94 mean F1-score. However, there exists potential for enhancements in the intra-section classification, which necessitates more focused feature extraction and model tuning tailored to the specific classes involved. Consequently, the results imply that a single model may face difficulties in sufficiently predicting all classes.

Table 6.6: Comparison of the performance of the XGBoost classifier with hand-crafted features and the *tsfresh* feature extraction method and an end-to-end deep learning-based ResNet over five-fold cross-validation. The best value of each metric is highlighted.

Setting	Feature	F1	ACC	PPV	TPR
low	hand-crafted	0.402 ± 0.034	0.407 ± 0.033	0.405 ± 0.034	0.407 ± 0.033
medium	hand-crafted	0.433 ± 0.024	0.435 ± 0.024	0.435 ± 0.023	0.436 ± 0.024
high	hand-crafted	0.535 ± 0.054	0.541 ± 0.051	0.551 ± 0.051	0.539 ± 0.051
low	tsfresh	0.434 ± 0.033	0.438 ± 0.031	0.442 ± 0.029	0.437 ± 0.031
medium	tsfresh	0.503 ± 0.043	0.506 ± 0.041	0.508 ± 0.044	0.505 ± 0.041
high	tsfresh	$\mathbf{0.641} \pm 0.065$	$\mathbf{0.644} \pm 0.063$	$\mathbf{0.653} \pm 0.064$	$\mathbf{0.642} \pm 0.062$
low	ResNet	0.294 ± 0.107	0.332 ± 0.077	0.328 ± 0.129	0.331 ± 0.077
medium	ResNet	0.442 ± 0.038	0.465 ± 0.038	0.489 ± 0.039	0.457 ± 0.038
high	ResNet	0.479 ± 0.054	0.492 ± 0.057	0.477 ± 0.052	0.493 ± 0.060

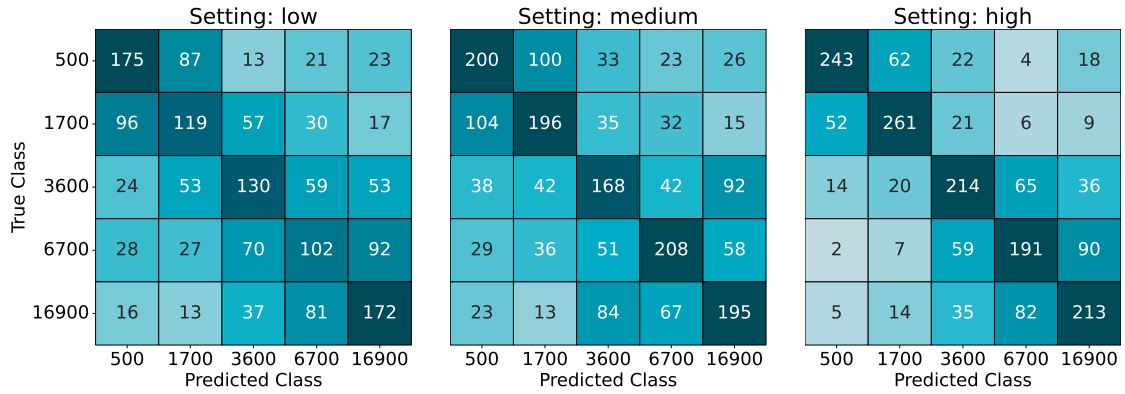


Figure 6.8: Summed up confusion matrix of the XGBoost classifier with the *tsfresh* feature extraction method over five-fold cross-validation. Misclassifications are mainly formed by two sections 500 to 1700 and 3600 to 16 900.

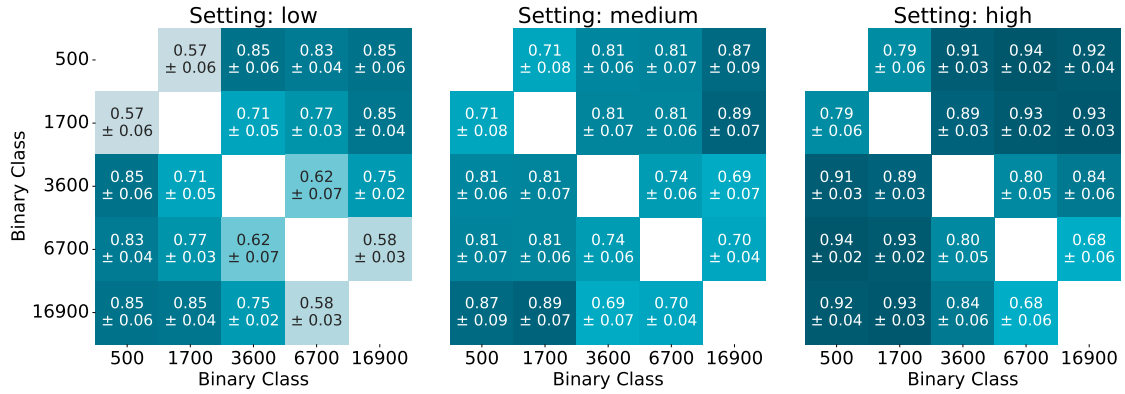


Figure 6.9: Results in F1 mean and standard deviation of the XGBoost classifier with the *tsfresh* feature extraction method for binary classifications. Figure is to be read by selecting a class from the y-axis and another from the x-axis to obtain the F1-score for its binary classification.

6.3.3 Hierarchical Classification[1]

Differentiation of multiple tissue stiffness classes is feasible to a certain extent, as shown in the previous section. It is observed that the classification is better the further away the classes are from each other in terms of stiffness. However, classifications that are close to one another are more difficult but not impossible as seen in binary classification. Thus, a single, flat classification model for all classes (as is the default in classical machine learning) is not sufficient to capture the complexity of the data and the classes. This motivates the usage of a hierarchical classification approach, which first classifies the

1 Parts of this section are based on the publication [51].

tissue stiffness into a broader category and then further classifies the tissue stiffness into a more specific category.

Data Preparation Tissue models from the data set D6 are used and are grouped according to Bin^1_{hier} and Bin^2_{hier} into these stiffness classes: *very soft*, *medium soft* and *firm*. The two classes *very soft* and *medium soft* are considered to belong to a parent category *soft*, thus reflecting a hierarchical structure of the data. Each level of the hierarchy is more refined in terms of tissue stiffness. Only data with a *high* generator setting is used, as it is observed that the classification performance is higher for this setting. Due to the highest performance in the previous section, feature extraction is performed on the data to obtain a feature vector for each contact. To this end, *tsfresh* is employed for automated feature extraction around the contact points of the time-series data. Feature extraction is conducted independently for every classification task within each hierarchy level, as well as for the flat classification task. Consequently, the choice of features varies depending on the specific classification objective. To address imbalanced data, all classes are randomly subsampled to the size of the smallest class, which happens to be the *very soft* class with 569 contacts. The data set is divided into two parts, one for training and one for testing. For each tissue stiffness class, 80 % of the data are allocated for training and 20 % for testing. Measures are taken to prevent any overlap between the training and test data, ensuring that data from the same tissue model is not used for both purposes. Consequently, each class has an equal number of 451 contacts for training and 118 contacts for testing. This ensures that the classification models are not biased towards the majority class and facilitates a clearer interpretation of the results.

Methods Classification of performed using a flat classification approach and a hierarchical classification approach, both visualized in Figure 6.10. Since both the flat and hierarchical classification approaches are agnostic to the classification model, any classification model can be used. In case of hierarchical classification, the classification models per parent node are trained independently and can therefore be different for each parent node. This approach allows using the classification model that is best suited for the data associated with the parent node.

Training is structured identically for each classification task (i.e. classifiers in Figure 6.10), regardless of classification approach or parent node. For each provided classification model, a grid search is performed to find the model's best hyperparameters. The grid search is performed by using a five-fold cross-validation on the training data with a stratification on the tissue models. After training all available classification models, the best performing model, along with its optimal hyperparameters, is automatically chosen as part of the workflow. This is done based on the weighted F1-score of the five-fold cross-validation. Afterwards, the best-performing classification model with its

corresponding hyperparameters is trained ten times with different initializations on the training data using bagging. This leads to ten saved trained models per classification task, all trained slightly differently due to the different initializations and bagging.

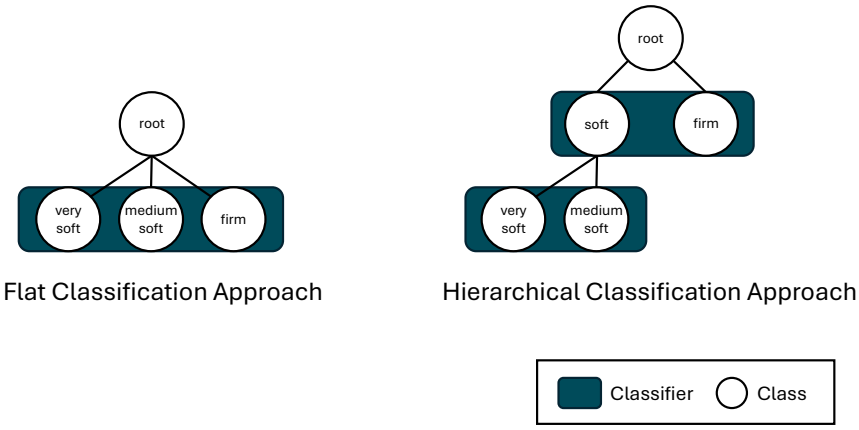


Figure 6.10: Flat and hierarchical classification approaches for the classification of tissue stiffness classes. White circles represent the classes in the hierarchy, blue boxes represent classifiers predicting the enclosed classes. Root indicates the root node of the hierarchy where the data enter the model – for comparison purposes, root is also used in the flat approach.

The ten saved trained models per classification task are used in an ensemble during testing, where the final prediction is the majority vote of the ten trained models. In the context of this work, the relative majority or agreement is considered as a certainty score, which is used to determine the confidence of the prediction. In other words, the certainty score is the percentage of the ten trained models that predicted the same class. Depending on the number of possible classes, the minimum certainty is 40 % for three classes (e.g. [1, 1, 1, 1, 2, 2, 2, 3, 3, 3]) and 60 % for two classes (e.g. [1, 1, 1, 1, 1, 1, 2, 2, 2, 2]). In case of three classes, there can be a tie between two classes, which is resolved by randomly selecting one of the two classes. In hierarchical classification, testing can be performed in a mandatory leaf node prediction (MLNP) or a non-mandatory leaf node prediction (NMLNP) approach. The MLNP approach requires the prediction to be assigned to a leaf node of the hierarchy. However, in the NMLNP approach, predictions are not restricted to the leaf nodes and can be assigned to intermediate nodes. When the certainty score is lower than a predetermined confidence threshold, the prediction is considered as uncertain. Within the NMLNP approach, the prediction is then assigned to the parent node of the predicted class, excluding instances where the parent node corresponds to the root. In such cases, or within the MLNP and flat approaches, the prediction is discarded and labeled as *uncertain*.

A variety of well-established classification models are used to evaluate the performance of the flat and hierarchical classification approaches. Since the classification task is a multi-class classification task, the classification models need to be able to perform multi-class classification. The following classification models are used in this work:

- Extreme Gradient Boosting (XGBoost)
- Support Vector Machine
- Logistic Regression
- k-Nearest Neighbors
- Gaussian Naive Bayes
- Linear Discriminant Analysis Classifier
- Quadratic Discriminant Analysis Classifier
- Perceptron

While all of these models are trained, it is important to note that the testing phase exclusively employs the automatically selected best performing models. The classification models are provided by the Python library scikit-learn [131] and XGBoost [30]. Additionally, all models are combined with a principal component analysis (PCA) as an additional feature reduction step which is optimized during the grid search. The PCA is used to reduce the dimensionality of the feature space and to remove redundant features to improve the performance of the classification models. Prior to that, a feature selection preprocessing based on the 10 % most important features according to an ANOVA F-value is performed to further limit the number of features. An extensive list of the hyperparameters that are optimized during the grid search can be found in the appendix in Table C.4.

Multiple metrics are used to evaluate the performance of the classification approaches. The metrics are calculated on the test data set. The following metrics are employed in this work: the balanced accuracy (BA), the weighted F1-score (F1), the weighted positive predictive value (PPV), the weighted true-positive rate (TPR). These balancing and weighting schemes are used to account for the class imbalance in the data at inference that can occur because of the possibility of discarding data due to low certainty. Nevertheless, these metrics are mostly designed for a flat classification approach and may lack the ability to evaluate the performance of a hierarchical classification approach. In order to incorporate the classification on different levels of the hierarchy, additional metrics are used to evaluate the performance of the hierarchical classification approach: the hierarchical positive predictive value (hPPV), the hierarchical true-positive rate (hTPR) and the hierarchical F1 (hF1) [96]. Details on these metrics can be found in section 3.3.

Table 6.7: Results of the flat and hierarchical classification approaches without enforced confidence thresholds on test data. The best value of each metric is highlighted.

Method	BA	F1	TPR	PPV	hF1	hPPV	hTPR
Flat	0.768	0.769	0.770	0.768	-	-	-
MLNP	**0.788**	**0.788**	**0.788**	**0.788**	**0.796**	**0.793**	**0.808**

Not Enforcing Confidence Thresholds Results of the flat and hierarchical classification approaches MLNP on test data without enforced confidence thresholds are shown in Table 6.7. In the absence of enforcing confidence thresholds, the predictions are determined by the majority vote of the ten trained models per classification task. No consideration is given to how many models agree on the same class and no predictions are discarded in this process. In case of flat classification, this means the minimum certainty score of 40 % is used, whereas for the hierarchical classification MLNP approach, the lowest certainty score of 50 % is used. The difference in the minimum certainty score is due to the fact that the hierarchical approach is structured in a binary tree in which the data are classified into two classes at each node. Higher values for all metrics are achieved by the hierarchical classification MLNP approach compared to the flat classification approach. The hierarchy-aware metrics hF1, hPPV and hTPR are only available for the hierarchical MLNP approach. The NMLNP approach is not evaluated in this section, since not enforcing confidence thresholds does not allow for the hierarchical classification to fall back to a parent node. These and all subsequent results are obtained by using the best performing classification model per classification task with its corresponding hyperparameters. The determination of the model is made based on its performance during the automated training workflow. In the flat classification approach, a Logistic Regression emerges as the optimized, best performing model. Meanwhile, in the hierarchical model, the root classifier is a Logistic Regression, and the child classifier is an XGboost tailored for the *soft* class. Optimized hyperparameters can be found in the appendix in Table C.5.

Enforcing Confidence Thresholds With the classifications being done by an ensemble of ten trained models per classification task, the certainty score, i.e. the ensemble agreement, can be used to determine the confidence of the prediction. This means that the predictions can be discarded if the certainty score is lower than a predetermined confidence threshold. The results for the flat and hierarchical classification approaches with enforced confidence thresholds are shown in Figure 6.11 – quantitative results can be found in the appendix in Table B.2. The confidence thresholds are varied from 40 % to 100 % in steps of 10 % for the flat classification approach and from 50 % to 100 % in steps of 10 % for the hierarchical classification approach – again, with the minimum value meaning that no data are discarded. Notice that the hierarchical classification NMLNP approach is not evaluated for a confidence threshold of 50 %, since this would

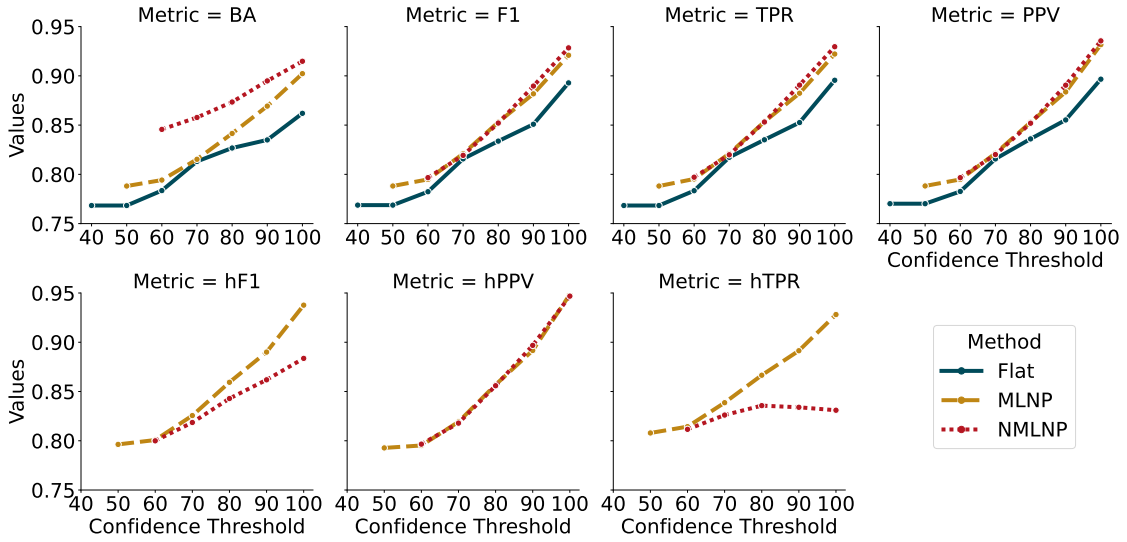


Figure 6.11: Results of the flat and hierarchical classification approaches with enforced confidence thresholds on test data.

result in the same results as the hierarchical classification MLNP approach. Data that are discarded due to low certainty are not considered for the calculation of the metrics. However, in case of the hierarchical classification NMLNP approach, data with a certainty score lower than the confidence threshold are set to the parent node of the predicted class, given that the parent node is not the root node. In this work, this results in an additional *soft* class for the hierarchical NMLNP approach, which is not present in the hierarchical MLNP approach. Generally, the results show an increase in the performance of the classification approaches with increasing confidence thresholds. Regarding the non-hierarchical metrics, both hierarchical approaches outperform the flat classification approach for almost all confidence thresholds. Notably, at the confidence threshold of 70 %, the same performance is achieved for the flat and hierarchical classification approaches. A higher performance can be observed for the hierarchical classification NMLNP approach compared to its MLNP counterpart. Especially in the BA metric, the difference is more pronounced. At 100 %, the NMLNP approach achieves a BA of 0.915, an F1 of 0.928, a PPV of 0.936, and a TPR of 0.930. The hierarchical metrics show lower results for the NMLNP approach compared to the MLNP approach. In particular, the hTPR for NMLNP is lower than the hTPR for MLNP for all confidence thresholds and is stagnating. This behavior is attributed to the generalization error, which is caused by not predicting the most specific class, i.e. predicting *soft* instead of *very soft* or *medium soft* [28]. With the MLNP approach, the generalization error is not present, since the prediction is forced onto the most specific class. This effect is accordingly affecting the hF1-score as well.

When examining the distribution of certainty scores for both the flat and hierarchical

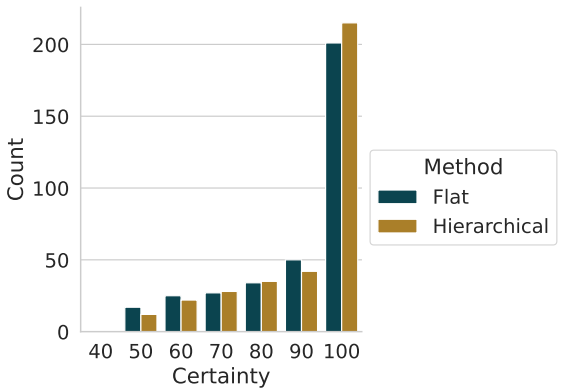


Figure 6.12: Distribution of the certainty scores for the flat and hierarchical classification approaches on test data.

classification approach, it can be observed that the majority of certainty scores reach 100 %, with this trend being slightly more pronounced in the hierarchical classification approach, as illustrated in Figure 6.12. It is important to emphasize that it is not necessary to distinguish between MLNP and NMLNP at this point, as this only affects the restructuring of the predictions and not the certainty scores themselves. Additionally, it is worth noting that the flat classification approach never attains the lowest possible certainty score of 40 %.

Setting the threshold for certainty scores involves a fine balance between optimizing performance and managing discarded data. To gain further insights into the impact of adjusting the confidence threshold, the absolute ratio of incorrectly classified data to correctly classified data among the discarded samples – both in relation to their initial amounts – can be examined. This ratio is calculated based on the differences between the confusion matrices obtained using a higher and a lower confidence threshold. In Figure 6.13, exemplary confusion matrices of the NMLNP approach are depicted within red boxes. These matrices demonstrate the resulting difference on the right from the utilization of a 70 % confidence threshold (lower confidence threshold) to a 80 % confidence threshold (higher confidence threshold). The number of incorrect classifications within this difference (Figure 6.13: off-diagonal entries in difference matrix) is divided by the sum of incorrect classifications in the confusion matrix associated with the lower confidence threshold (Figure 6.13: off-diagonal entries in red box of 70 % matrix). Similarly, the number of correct classifications within the difference (Figure 6.13: main diagonal entries in difference matrix) is divided by the sum of correct classifications in the confusion matrix associated with the lower confidence threshold (Figure 6.13: main diagonal entries in red box of 70 % matrix).

The resulting ratio between these two relative values is then calculated and presented in Figure 6.14. When the ratio is less than 1, increasing the confidence threshold leads to a reduction in performance, as relatively more correct classifications are discarded

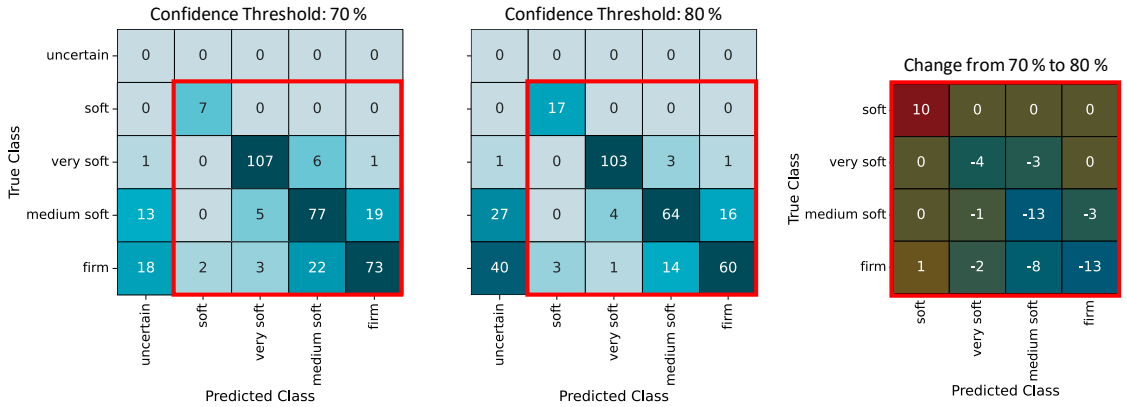


Figure 6.13: Exemplary confusion matrices of the NMLNP approach for confidence thresholds of 70 % and 80 % on left and middle, respectively. The difference between the two confusion matrices is shown on the right. Discarded data are shown in virtual *uncertain* class outside the red box. *soft* class predictions are caused by the hierarchical NMLNP approach, which allows falling back to a parent node. This allows to recover otherwise discarded data as *soft* class, indicated by the positive sign in the difference matrix on the right.

compared to incorrect ones. Conversely, when the ratio exceeds 1, increasing the confidence threshold enhances performance, even though some data are discarded in the process. The higher the ratio, the better the performance is enhanced by increasing the confidence threshold. The results indicate that the NMLNP approach almost consistently outperforms both the MLNP and flat classification approaches. Overall, all ratios exceed 1, implying that elevating the confidence threshold consistently yields performance enhancements. When examining the ratios between consecutive confidence threshold, the NMLNP approach exhibits values above 3.53. Its highest value of 5.08 is observed from 70 % to 80 % (see Figure 6.13) and is mainly due to the recovery of otherwise discarded data as a *soft* class during the process, as indicated by the positive sign. In contrast, the maximum value for the MLNP approach is 3.63 when moving from 60 % to 70 %, while the maximum value for the flat classification approach is 6.43 transitioning from 60 % to 70 % as well. Similarly, when assessing the ratios between the lowest and subsequent ones, consistent trends can be observed across the respective approaches. Generally, a higher overall performance across all confidence transitions can be observed for the NMLNP approach. In the case of the flat and NMLNP approaches, robust performance is especially evident within the lower threshold ranges, up to 70 %, after which the performance begins to stagnate.

Effect of Discarded Data Only considering the data that are not discarded due to low certainty can lead to a bias in the evaluation of the classification approaches. This is due to the fact that the discarded data are not directly considered for the calculation of

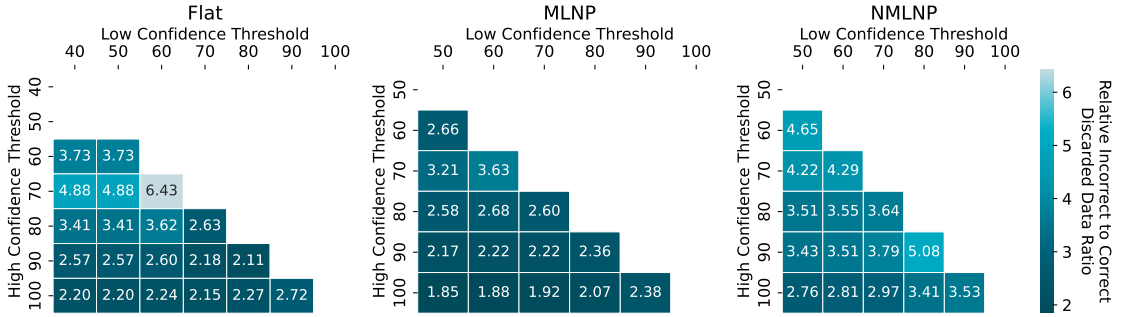
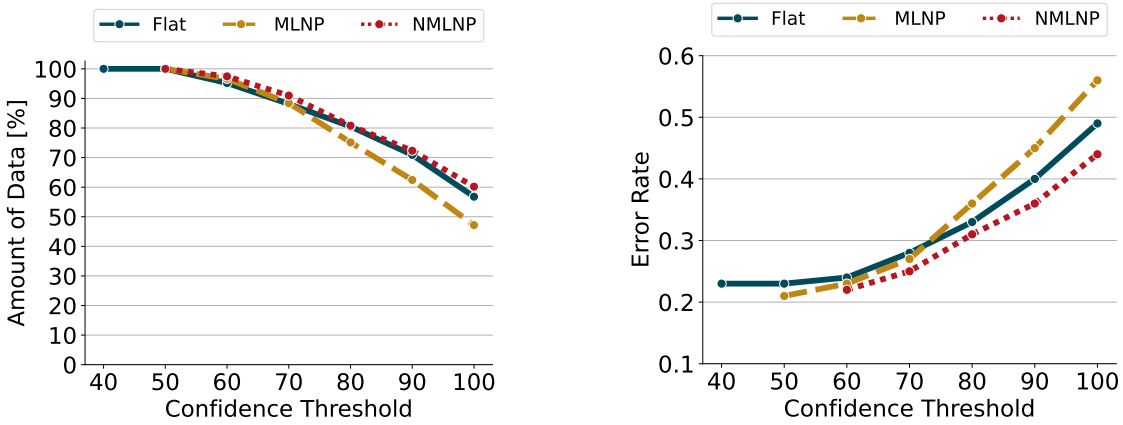


Figure 6.14: Absolute ratio of relative incorrect to correct classified data that are discarded due to low certainty. The ratio is calculated based on the differences between the confusion matrices obtained using a higher and a lower confidence threshold. Figure is to be read from Low Confidence Threshold to High Confidence Threshold. Values larger than 1 indicate that the performance is enhanced by increasing the confidence threshold, while values smaller than 1 indicate that the performance is reduced.

the metrics. As already stated, the aim should be to remove as little data as possible, while at the same time achieving high performance when increasing the confidence threshold. The effect of discarding data on the amount of initially available data is shown in Figure 6.15(a). The amount of data are shown as a percentage of the initially available data. As anticipated, the volume of retained data diminishes with increasing confidence thresholds, ultimately reaching its lowest point at a 100 % confidence threshold. In the



(a) Percentage of data that are not discarded

(b) Error rate with *uncertain* data

Figure 6.15: Percentage of data and error rate for flat and hierarchical classification approaches. The percentage of data are shown as a percentage of the initially available data. The error rate is calculated as the ratio of false predictions (including those marked as *uncertain*) to all predictions.

end, the MLNP, flat and NMLNP approaches preserve 47.2 %, 56.8 %, and 60.2 % of the original data, respectively. In relative terms, the NMLNP approach discards the smallest portion of data, followed by the flat approach and the MLNP approach which discards the most data. A comparable trend between the NMNLP and flat approach can be noted in regard to the relative amount of discarded data.

To better assess the discarding effect, a virtual *uncertain* class is introduced, encompassing all data that get discarded due to low certainty. Exemplary confusion matrices for the NMLNP approach are presented in Figure 6.13 for 70 % and 80 % confidence thresholds with the virtual *uncertain* class located outside the red boxes. An error rate is then calculated for each confidence threshold, defined as the ratio of false predictions (including those marked as *uncertain*) to all predictions (Figure 6.13: off-diagonal entries to all entries). It is important to note that the absolute values of the error rates should not be evaluated on their own. Instead, they should be compared relative to each other, considering that *uncertain* is a virtual class. The error rate is shown in Figure 6.15(b). It can be seen that across all methods the error rate increases with increasing confidence thresholds. It is noticeable that the error rate for the hierarchical classification NMLNP approach is lower than the error rate for the other approaches. This effect can be attributed to the fact that the hierarchical classification NMLNP approach is able to fall back to a parent node, thus reducing the number of false predictions. Therefore, as evident in Figure 6.15(a), fewer data are discarded, which results in an overall lower error rate. However, it can be seen that in the beginning similar error rates are achieved for all approaches, which start to diverge with increasing confidence thresholds starting from 70 %.

6.3.4 Contextualization and Synthesis of Cumulative Findings

Tissue differentiation with an ultrasonic aspirator under realistic conditions in a handheld mode proves to be a challenging task, as the results showed. However, important insights could be gained from the results of the classification approaches.

In an initial experiment, the handheld resection process was carried out on a reduced data set to investigate the differentiation dependency on the generator setting and the overall capability of tissue differentiation. The results clearly indicated influence of the generator setting, showing that leveraging the generator setting as a prior information can enhance the classification performance. By training a classification model for each generator setting, an overlap of features was reduced, and the learning process was eased. Furthermore, the results showed that classification in a handheld mode is feasible. Nevertheless, the results were generated on a reduced data set with a limited variety of tissue models per class and classes in general.

Thus, in a second experiment the classification performance was evaluated on a more diverse and larger data set with five classes and three generator settings *low*, *medium* and *high*. Feature extraction was performed – as in previous experiments – with hand-

crafted features and with a dedicated feature extraction method for time-series data, *tsfresh*. In addition, predictions were made per contact rather than for each temporal data point at a sample-based level. Furthermore, feature extraction was conducted with a deep learning approach, which worked in an end-to-end manner and was based on a ResNet architecture. Performance showed that the *tsfresh* feature extraction method outperformed both the hand-crafted features and the deep learning-based approach, especially for the *high* generator setting. Even though deep learning excels in various applications, it is worth noting that substantial time and hardware resources are not always essential for achieving satisfactory results with the specific data set at hand. In fact, utilizing hand-crafted features proved to yield comparable outcomes. Furthermore, opting for alternative feature methods such as *tsfresh* and straightforward classification algorithms led to even better performances in the context of this particular data set. The consistent superiority of the *high* generator setting in achieving optimal performance across various feature extraction methods suggested a notable dependency on the generator setting. One hypothesis that may be considered is that this dependency is a consequence of the level of control that the user has over potential movements and speeds during tissue resection. In the event that a higher generator setting is applied, the rate of tissue removal may be accelerated, thereby reducing the likelihood of external factors, such as movements or speed fluctuations, influencing the procedure. This would result in more reliable and consistent data, thereby enhancing performance across feature extraction methods. In contrast, lower generator settings would entail slower tissue resection, which could potentially result in greater susceptibility to external influences. Factors such as movements or variations in speed may interfere with the resection process, leading to less consistent data and inferior performance. While this hypothesis could explain the observed behavior, its confirmation would likely require specifically designed studies to directly assess the impact of movement and speed variability on tissue resection outcomes across different generator settings.

Misclassifications were observed mainly between certain sections of the confusion matrix, indicating that measurements are difficult to classify when they are close to one another in terms of tissue stiffness under the scope of the used features and classification models. Thus, a stiffness distance dependency could be observed. By relaxing the classification to an isolated binary classification task, performance was improved and had the potential to better differentiate between the classes. This encouraged the adoption of a hierarchical classification approach, initially categorizing tissue stiffness into broader classifications before delving into more specific classes.

To this end, the third experiment was conducted to evaluate the performance of a hierarchical classification approach against a flat classification approach on a *high* generator setting. The former was further divided into two approaches, namely the MLNP and NMLNP approaches, which differ in their ability to fall back to a parent node while classifying new data. Overall, the results showed that a hierarchical NMLNP approach outperforms both the flat and MLNP approach with metrics such as F1 reaching

up to 0.92 and a lower amount of discarded data. This outcome can be attributed to the NMLNP approach's capability to fall back to a more general class when the certainty of a prediction is lower than a predetermined confidence threshold. On the one hand, this makes the performance superior to a classical flat classification approach and on the other hand it allows for a more comprehensible solution to the outcome. This is especially important in the context of machine learning and clinical application, as it allows explaining the classification results to the surgeon and to provide a basis for the surgeon to make a decision. Additionally, this can lead to more acceptance of machine learning-based methods in the medical field. Evidently, the hierarchical classification model used in this work is only a proof of concept and can be extended to a multi-level hierarchy, where the data classification gets even more specific with each level of the hierarchy. This, however, requires a larger and more diverse data set in terms of tissue stiffness, which will be addressed in future work.

The results also showed that the performance of the classification approaches can be enhanced by increasing the confidence threshold. Even though this consistently led to better and more robust performance for all approaches, the amount of data that is discarded in this process needs to be considered. In challenging scenarios, the potential dismissal of every other classification in a realistic setting could frustrate surgeons and hinder the acceptance of this method. Consequently, the confidence threshold needs to be chosen carefully and a trade-off between performance and discarded data needs to be found. Within this work, an 80 % confidence threshold appeared to offer an effective balance, achieving high performance exceeding 0.8 in various metrics while preserving over 80 % of the data, especially in the hierarchical NMLNP approach. These trade-offs can be further explored in future research by expanding the ensemble size, currently set at ten trained models per classification task and by employing a variety of classification models within the ensemble.

6.4 Impact of Force on Tissue Differentiation

One factor which might vary between users and environments such as laboratory and clinic is the force applied to the tissue. Controlling the force is a subjective task and bound to the user's experience. Additionally, keeping the force constant is a challenging task, especially in a clinical environment where a surgeon has to focus on the resection process, while standing and moving in an uncomfortable position. Hence, the question arises whether the force applied to the tissue has an impact on tissue differentiation.

To investigate this, force measurements are conducted alongside the resection process in the laboratory environment and are given in data set D7. For this study, the generator setting *high* is used, as it exhibits the best performance in tissue differentiation. Four tissue models belonging to the classes *soft* and *firm* are used for the measurements.

In Figure 6.16, the distribution of the applied forces to the two tissue classes is depicted. Two distinct distributions can be observed, with the *soft* tissue models exhibiting a lower

force distribution compared to the *firm* tissue models. Even though the user was asked to apply forces higher than usual, it can be observed that for the *soft* tissue models a limit is reached which does not exceed 50 mN. For *firm* tissue models, the force applied is higher and reaches up to 175 mN. This shows that the tissue stiffness has an impact on the force applied to the tissue, which is to be expected as a *soft* tissue model is exhibiting such a stiffness that high forces are pushing through the tissue, making it impossible to apply higher forces at all. The *firm* tissue models are more resistant to the force applied, which allows for higher forces to be applied to the tissue. In terms of relevance, those findings are in line with internal studies showing that neurosurgeons are applying average forces of 30 mN and with the majority of human brain tissue stiffness values being in the *soft* tissue range, c.f. Figure 6.18.

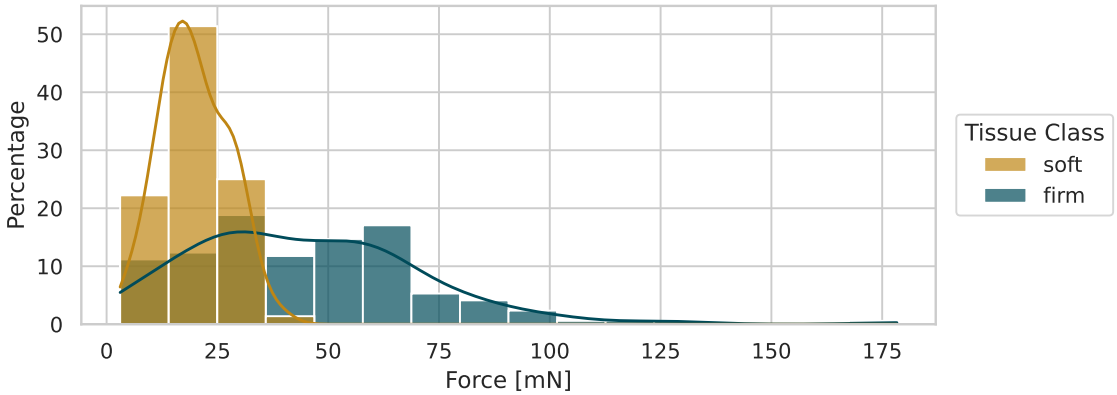


Figure 6.16: Distribution of the applied forces to the tissue models of the classes *soft* and *firm*.

However, what is the impact of the force applied to the tissue on the tissue differentiation? The data are processed with the *tsfresh* feature extraction method and classified with the NMLNP approach, only using the first level of the hierarchy for *soft* and *firm* classification without a certainty threshold. Figure 6.17 shows the common feature with the highest Spearman correlation coefficient per tissue class over the applied force. The 25 highest correlated features are shown in the appendix in Table B.4 for each class. It can be observed that for the two tissue classes *soft* and *firm*, a low to moderate correlation of 0.273 and -0.449 is present, respectively. However, as visible in Figure 6.17, though a correlation is present, the classification performance is not directly influenced by the force applied to the tissue. Even though misclassifications are present, no clear trend can be observed in the classification performance over the applied force. This is further supported with a Chi-squared test, which shows no significant difference in the correct classification performance over the applied force ($p > 0.05$). Though not a clear trend, some misclassifications can be found for the *firm* tissue model in the higher force range over 50 mN. This could be attributed to the fact that this force range is not commonly encountered in the training data, given that the procedure is conducted with

low forces, mirroring the clinical environment. Based on the given data and results, it can be concluded that features extracted indeed can be influenced by the force applied to the tissue, but the classification performance is robust against the force applied to the tissue within adequate force ranges.

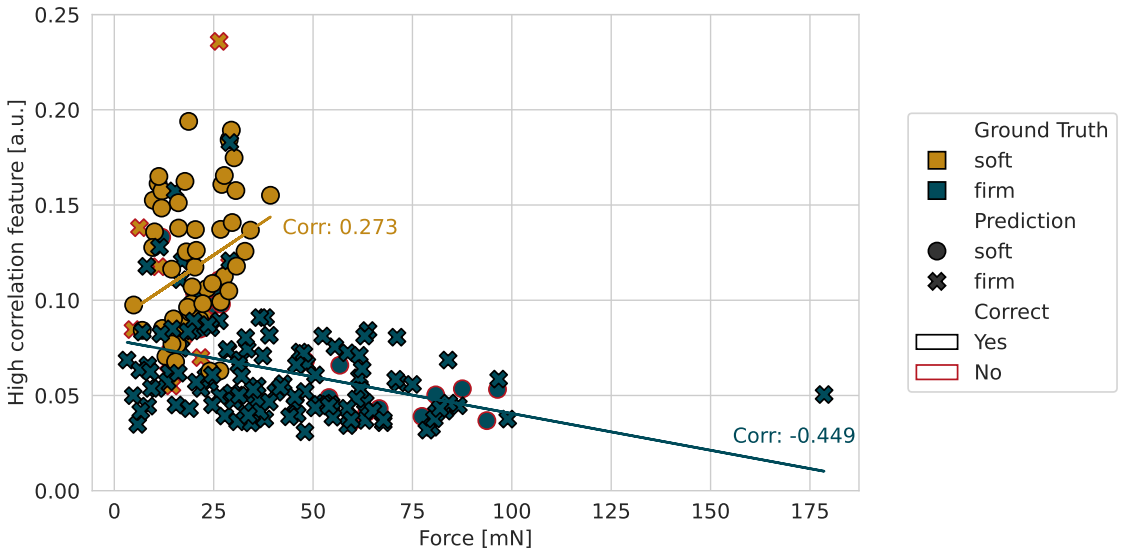


Figure 6.17: Classification of *soft* and *firm* tissue models on feature with the highest Spearman correlation per tissue class over the applied force. Correct predictions are shown with black edges, while misclassifications are shown with red edges.

6.5 Insights into Clinical Data

To ensure adequate data volumes for employing data-driven methods in tissue differentiation, the preceding results were obtained using artificial tissue models. Do these laboratory-based findings translate to the clinical setting? At first, the stiffness of the tissue models is compared to the stiffness of the tissue samples from the clinical setting, to assess the similarity of the data. In Figure 6.18, the Young's modulus in the laboratory environment is compared to the Young's modulus of the tissue samples from the clinical environment. It can be observed that the majority of tissue samples of both environments are located within the same Young's modulus range. This indicates that the tissue models used in the laboratory environment are a good representation of the tissue samples found in a typical clinical environment. An exact matching representation is not to be expected, as the process of artificial tissue model creation is subject to fluctuations, making the process of targeted tissue creation difficult. Furthermore, one has to consider that the clinical tissue samples are limited by the availability and changing conditions during surgery, of which only 26 samples are available for the analysis. The red vertical lines in Figure 6.18 represent the hierarchical bin edges. It is evident from the

figure that the choice of hierarchical classification bin edges aligns well with the clinical data. This observation suggests that adopting a hierarchical classification approach is viable within a clinical context.

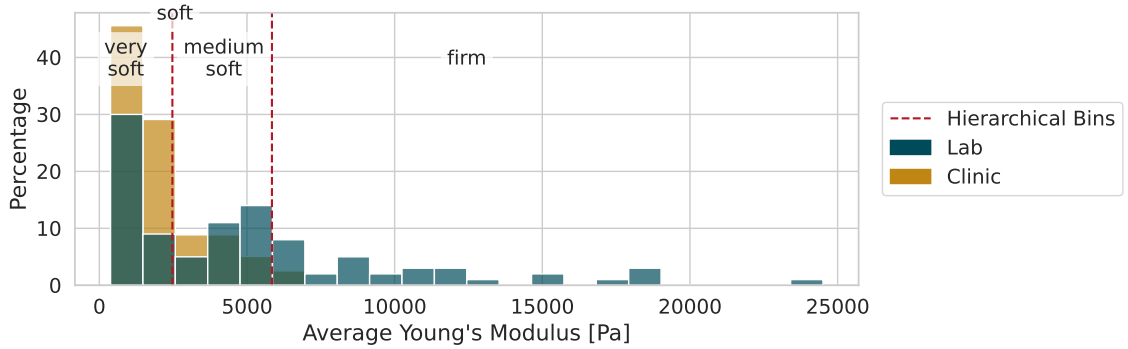


Figure 6.18: Comparison of the Young's modulus of the tissue models in the laboratory environment and the tissue samples from the clinical environment. Red vertical lines indicate the bin edges of the hierarchical classification approach.

Is there a distinction between non-expert users and expert users, which could lead to a user dependency bias? To answer this, an expert (neurosurgeon, more than 20 years of experience) is asked to resect a tissue model in a laboratory environment with the Gen_{dev} generator. The data of the expert user are compared to the data of a non-expert user in the laboratory environment from D6 based on the same batch ID and settings. The findings, illustrated in Figure 6.19, outline the relative voltage shift. Upon comparison, it becomes apparent that both users exhibit similar distributions, with the expert user demonstrating a slightly narrower spread yet a comparable median. This coherence in the data suggests a lack of, or at least minimal, user dependency bias. Nevertheless, for a definitive conclusion regarding user dependency, additional data from various users, as well as a more diverse data set concerning tissue stiffness, is imperative.

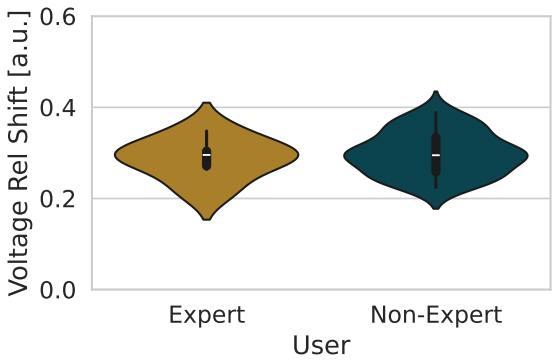


Figure 6.19: Comparison of the relative voltage shift of the tissue models in the laboratory environment for expert and non-expert users.

Comparing data acquired during neurosurgical procedures with data from laboratory-based experiments is only possible to a limited extent. This is mainly due to three reasons: First, electrical data from the ultrasonic aspirator can not be directly associated with the tissue stiffness, because of the acquisition process. It is not guaranteed that the electrical data are acquired on the same tissue as tissue for the stiffness measurement. Hence, a relationship on the basis of tissue stiffness would be difficult to establish and could lead to misleading results. Second, the settings of the generator are dependent on experience and subjective assessment of the surgeon, leading to a high variability in the generator setting. This makes a direct comparison between the laboratory and clinical data possible to a certain extent, due to the fact that a dependence on the classification performance on the generator setting is observed. An overview of clinically used settings in comparison to the laboratory settings is shown in Figure 6.20. Even though higher settings do not necessarily lead to faster or more efficient tissue resection, a slight tendency towards using higher settings is observed in the clinical environment.

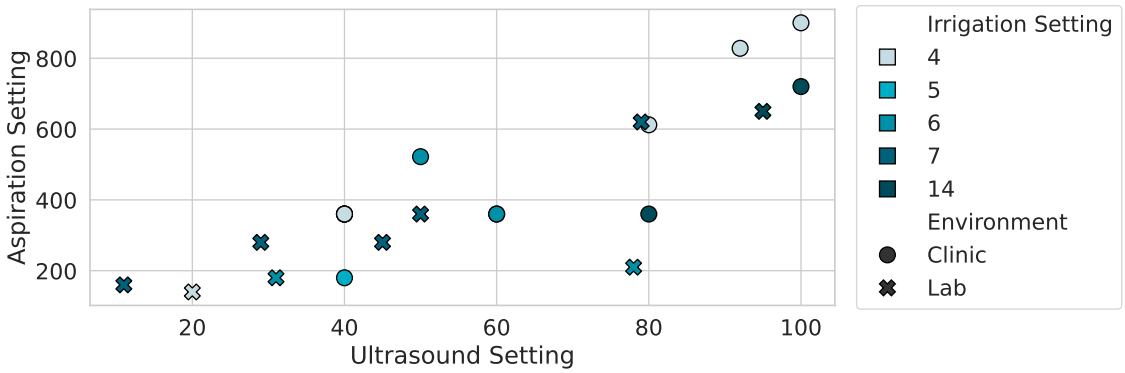


Figure 6.20: Comparison of the generator settings used in the laboratory environment and the clinical environment.

Third, the generator Gen_{prod} used in the clinical setting is not the same as the generator Gen_{dev} used in the laboratory setting for the most advanced results. This is because, at the time of this writing, the generator Gen_{dev} is in a development phase and not yet approved for clinical use. One of the main differences between the two generators is the sampling frequency of 21 Hz for the clinical generator and 1000 Hz for the laboratory generator. This considerable difference in the sampling frequency has an impact on the available data and the feature extraction process. In other words, for feature extraction in the laboratory environment, a *contact* phase of about 500 data points is expected, whereas in the clinical environment, fewer than 11 data points are available. Often even fewer data points are available, due to too short *contact* phases established by the surgeons in the asked ten individual contact cases at the beginning of the usage of the ultrasonic aspirator. Thus, feature extraction similar to the laboratory process proves to be infeasible.

However, with limited possibility to classification on clinical data from data set D8, the legitimacy of simple features such as the relative voltage shift can be compared to the laboratory data. To this end, data from D4 is used, which is acquired using the same generator type Gen_{prod} as in the clinical environment. A comparison between the relative voltage shift for both, the laboratory and clinical data are shown in Figure 6.21. It can be seen that a similar range in distribution of the relative voltage shift is present in both environments and that the clinical data are not occupying a different range than the laboratory data. This suggests that data acquired in a laboratory environment are closely related to the data acquired in a clinical environment. This indicates that outcomes derived within the confines of the laboratory setting possess a degree of validity. Though no direct distinction can be made on basis of tissue stiffness, a distinction of the simple relative voltage shift feature can be given on the tumor entity encountered during the surgery. A depiction of the relative voltage shift for the different tumor entities is shown in Figure 6.22. Between the different tumor entities, a slight difference of the median and its distribution can be observed. However, the distinction is not clearly distinguishable and a clear separation of the tumor entities cannot be made given the current data. A more diverse and larger data set is needed to make a clear distinction between the tumor entities. Additionally, data on healthy tissue need to be included to enable a clear distinction between healthy and tumor tissue. Despite the limitations in directly comparing clinical and laboratory-generated data, the findings and insights derived from laboratory data possess a relevance that may facilitate their translatability to clinical settings in the future. A proper validation of data and classification models is of utmost importance and is yet to be conducted in a clinical setting to ensure the generalizability and reliable transferability of the results using the Gen_{dev} generator.

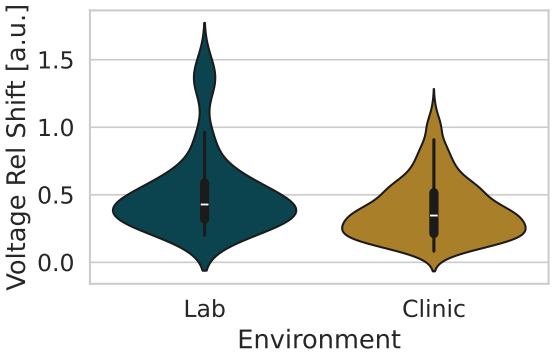


Figure 6.21: Comparison of the relative voltage shift of the tissue models in the laboratory environment and the tissue samples from the clinical environment.

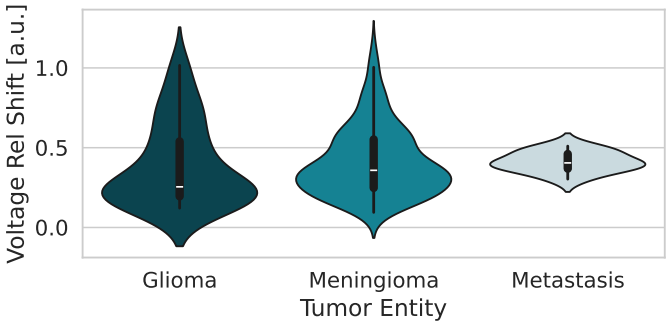


Figure 6.22: Comparison of the relative voltage shift for different tumor entities.

6.6 Assessing Human Tissue Differentiation Capability

Tissue differentiation with an ultrasonic aspirator seems feasible under realistic conditions in a handheld mode. The results show high performance in the classification of tissue stiffness classes, especially for the *high* generator setting. This, however, begs the question of how good is this performance in relation to a human's capability? For this, a human performance comparison study is conducted in which participants are asked to classify tissue models in a blinded condition after an initial training phase in the classes *soft*, *medium*, and *firm*. These classes reflect exactly the previously used classes *very soft*, *medium soft*, and *firm* and have been adapted to make them easier to understand for the participants. Participants' classifications are collected in the data set D9 throughout the study.

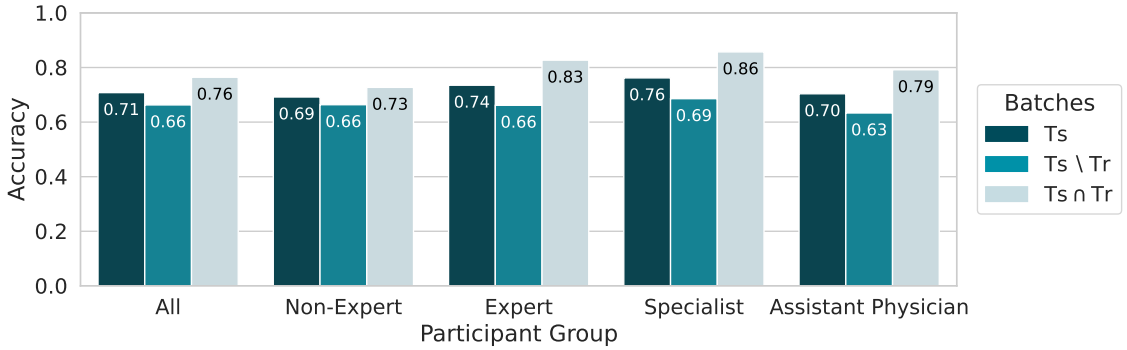


Figure 6.23: Results of the human performance comparison study with the distinction of different participant groups and batch IDs. Note that Non-Expert and Expert are subgroups of the participant group All, while Specialist and Assistant Physician are subgroups of the Expert group. Ts represent all test batch IDs. Ts ∩ Tr represent the test batch IDs that are also present in the training data. Ts \ Tr represent the test batch IDs that are not present in the training data.

Results of the human performance comparison study are shown in Figure 6.23, detailed quantitative results can be found in Table B.3. The results are provided for all batch

IDs belonging to the test data (Ts), as well as for solely batch IDs that are present during training (Tr) as in Ts ∩ Tr and those that are not present during training as in Ts \ Tr. By this, effects of the training data on the performance can be assessed. It is important to note that tissue models from the same batch ID are not identical, just that they are from the same production batch. Observing the results without differentiating between different user groups (All) shows the highest performance for Ts ∩ Tr with an accuracy of 0.76. A difference in performance of five percentage points between Ts and the subset Ts ∩ Tr is present, while this difference is doubled between Ts \ Tr and Ts ∩ Tr. Both show no significant difference in performance between the two groups ($p > 0.05$), indicating no training data bias in human performance under the given conditions. Nevertheless, in ideal conditions, the training phase would consist of a high variety of tissue models, which could allow for a clearer understanding of class properties. However, in case of human studies, a vast and exhaustive training phase is not feasible under limited data availability and manufacturing constraints. Still, humans are given the benefit of having a general understanding of stiffness and tactile feedback, making it easier to learn from limited data, as opposed to a machine learning model.

Examining the user groups from the perspective of expertise (neurosurgeons) or no expertise shows an overall higher performance for the expert group. However, there is no significant difference ($p > 0.05$) in performance between the two groups in any batch configuration. Further differentiation can be made within the expert group by distinguishing between specialist and assistant physicians to determine if experience level has an impact on performance. Despite the higher performance observed for the specialist group, no significant difference in performance between the two groups is observed for any of the batch constellations ($p > 0.05$).

Independent of user groups and batch considerations, the performances show that misclassifications are present. A detailed examination of the confusion matrix using all participants and all batch IDs is presented in Figure 6.24.

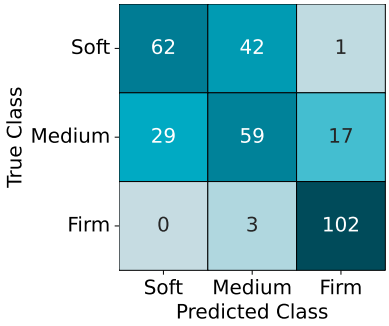


Figure 6.24: Confusion matrix of the human performance comparison study. Challenges in distinguishing between the *soft* and *medium* classes are apparent, contrasting with strong performance in identifying the *firm* class.

It is evident that the *firm* class exhibits exceptional classification accuracy, whereas participants encounter challenges in distinguishing between the *soft* and *medium* classes.

Which specific tissue models are difficult to classify can be seen in Figure 6.25. The figure shows how often the tissue models (shown by their tissue ID) are classified correctly by the participants as a ratio of the total amount of classifications of the tissue model. The tissue ID is based on the batch ID appended by the number of the tissue model of that batch. Evidently, from the *firm* tissue models, only one tissue model is not classified entirely correct, which happens to be the *firm* tissue model with the lowest Young's modulus. The results on tissue IDs M-B-1 and M-B-2 display an interesting behavior, as they are from the same batch ID but show different correct classification ratios. A similar behavior can be observed for tissue IDs S-B-1 and S-B-2. This can be attributed to the fact that the tissue models are not necessarily identical, even though they are from the same batch ID. Nevertheless, a higher performance would be expected for the tissue models from the batch ID M-B since the participants were trained on the same batch ID. This indicates the difficulty of the tissue differentiation for humans, especially within the softer stiffness domain. Differentiating between very soft and very firm tissues seems to be easier for the participants, as indicated by the correct classification ratio for the tissue models with Young's modulus of lower than 1000 Pa (S-A-1) and higher than 7000 Pa (all *firm* tissue models).

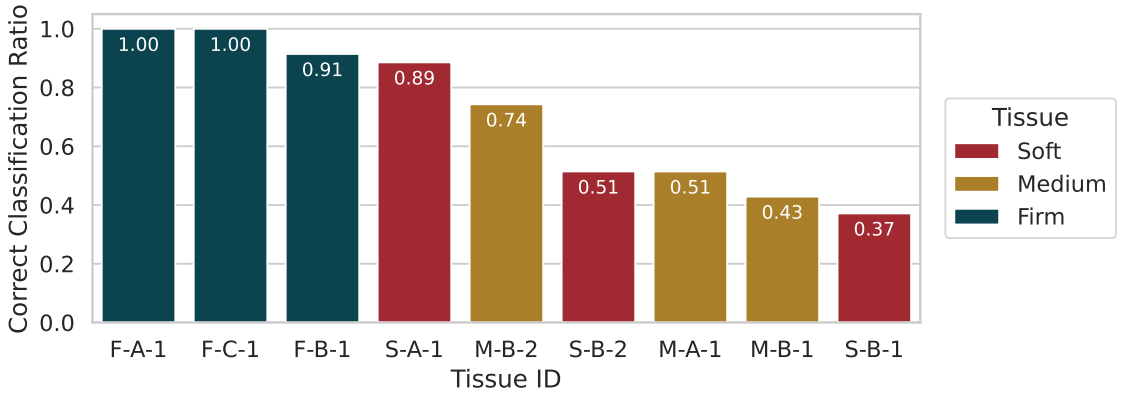


Figure 6.25: Correct classification ratio per tissue model for the human performance comparison study. Correct classification ratio is shown as a ratio of the total amount of classifications of the tissue model. The tissue ID is based on the batch ID appended by the number of the tissue model of that batch.

Considering the results of the human performance, e.g. F1 of 0.668 to 0.765 for all test batch IDs (c.f. Table B.3) in relation to the machine performance with an F1 of 0.788 using MLNP (c.f. Table 6.7), it can be observed that the machine performance is on par or even better than the human performance. Even though the data sets considered in the human performance comparison study and the machine performance evaluation are not

identical, a comparison is still valid given the variety and vast amount of employed tissue models in machine performance evaluation. This suggests that a machine performance is comparable to the human performance in terms of tissue differentiation capability and even outperforms the human performance in some cases. Furthermore, this shows that the simplistic setup of the hierarchical classification approach is already a viable method for tissue differentiation, as human participants are not able to differentiate between the tissue models with a higher performance than the machine learning models.

6.7 Summary

In this chapter the feasibility of differentiating between different tissues using an ultrasonic aspirator is investigated. Tissue differentiation within the context of this work is limited to their stiffness and based on the Young's modulus of the provided and acquired tissue models and tissue samples and different labeling processes. Even though this is only a small part of possible tissue properties, it can be seen as an important part of telling apart different tissue types and is a first step in the novel field of tissue differentiation on the basis of a working ultrasonic aspirator. Tissue differentiation is seen in the overall context of the intelligent ultrasonic aspirator as the key component of an intelligent surgical system. This approach differs from previous techniques in that it simultaneously ablates tissue and determines its stiffness based on electrical data. This is a unique approach, particularly in the context of soft tissue and highly sensitive tissue found in neurosurgery. A comprehensive understanding of an intelligent system on which tissue it is working on is crucial for subsequent systems to make informed decisions and to assist the surgeon in the best possible way. Moreover, an uncertainty on a tissue stiffness prediction is clearly communicated, which is an important aspect in the acceptance of machine learning based methods in the medical field. With this work being focused on the feasibility of tissue differentiation in terms of stiffness, direct subsequent tasks are out of the scope of this work and thus not evaluated. However, a multitude of tasks can be envisioned. For example, a surgical margin detection system as seen in the work regarding the iKnife on skin and breast cancer [152, 80, 82, 153, 154] can be thought of, where the intelligent ultrasonic aspirator is used to detect the margin of a tumor and to communicate this to the surgeon. This means that an even larger and more clinically based study with a variety of tumorous and healthy tissues needs to be conducted, which is a challenging but essential and important task. This could not be adequately analyzed within the scope of this work, as the data are limited to a small variety of tissue samples. Other tasks could be directed towards the intelligent ultrasonic aspirator itself, such as the improvement and optimization of the resection process based on the predicted tissue stiffness. This would eliminate the user-dependent settings of the generator and could lead to a more efficient and faster resection process. A similar idea for ultrasonic surgical units was envisioned in [185, 184], where the power control of the device is dependent on the tissue type. However,

no further steps were taken to evaluate the feasibility of this approach.

This chapter evaluated the tissue differentiation capabilities of an ultrasonic aspirator with increasing complexity at each step. In the first step, it was initially examined whether there is any information at all in the data provided by the generator in order to realize a stiffness identification in a simplified case. The simplified case entailed the utilization of an alternative sonotrode type, in combination with the deactivation of the resection function and using a CNC machine. This showcased that differentiation between four different tissue models is possible resulting in a high performance. With this positive result, the next step was to evaluate the differentiation capabilities under the scope of tissue ablation, while still utilizing the CNC machine in a controlled environment. Taking into account the more difficult conditions, a differentiation of two artificial tissue types was successfully achieved with an F1-score of 0.9. This revealed that the influence of the generator settings, namely the irrigation, can cause certain signal characteristics that hinder the differentiation process. Filtering methods were applied to mitigate this influence and to improve the classification performance. Furthermore, no advantage of including temporal information in the classification process was observed – with the limitation of only using data during the *contact* phase. In the third step, the complexity was increased to a more clinical and realistic scenario, where a handheld resection process was conducted. This revealed that generator-dependent classification models can improve the classification performance, as opposed to a single model responsible for all generator settings. This can be seen as an implicit prior information, which can be used to enhance the classification performance. Furthermore, feature generation was moved to the specific point of contact state change from *no contact* to *contact* and its surroundings to leverage the most information possible with different feature generation methods. This showed the advantage of predefined and common features over deep learning-based feature generation methods, which most likely were limited by data availability and the specific task at hand. Finally, the hierarchical classification approach was introduced to improve the classification performance and to provide a more comprehensible solution to the outcome with simple classification models. This approach showed a superior performance in comparison to a classical flat classification approach with a performance of up to 0.93 and a lower amount of discarded data. Generally, it was observed that a high setting showed advantages in the classification performance over lower settings. Furthermore, the tissue stiffness was assigned into three categories: *soft*, *medium*, and *firm*, which corresponds to specific bins of the Young's modulus. Setting the results in perspective to the human performance comparison study, it was observed that the machine performance is comparable to the human performance and even outperforms the human performance in some cases. This legitimizes the use of machine learning models for tissue differentiation and shows the potential of the intelligent ultrasonic aspirator in the field of neurosurgery.

Aiming at potential future advancements, different avenues of exploration are conceivable. As seen in the work regarding the iKnife, mass spectrometry data can be used

as a reliable source of information for tissue differentiation. Interestingly enough, a similar approach by [4, 155] was already conducted on tissue waste from the ultrasonic aspirator. Combining mass spectrometry and the electrical data from the ultrasonic aspirator could lead to a more precise and reliable tissue differentiation as a multi-modal approach. This could even be further extended by integrating preoperative MRI images, which are navigated at the surgical site, to enhance the classification performance.

While the findings show promise, there remain limitations to address. Given that the findings predominantly rely on artificial tissue models, the extent to which they can be generalized to real-time tissue differentiation during surgery is somewhat restricted. Discrepancies in generators and settings between laboratory and clinical environments, coupled with limited clinical data availability, create challenges in directly comparing results. Nevertheless, there are overlapping characteristics in the data from both settings, suggesting potential for result transferability. Moreover, slight variations in features across different tumor types hint at the possibility of distinguishing between them. Yet, achieving a clear delineation between tumor types, as well as healthy tissue, requires a more diverse and expansive data set under the usage of the Gen_{dev} generator. Investigations into external influences, such as applied force on tissue, reveal no discernible impact on classification performance for distinguishing between *soft* and *firm* tissue models. Other potentially user-dependent factors, like instrument angle, movement direction or speed, were not explicitly examined but are implicitly present in the data due to multiple users and data acquisitions over the time of this thesis. Although explicit, verifiable links to such behaviors were not observed during data analysis, it is important to recognize their potential impact on classification performance, prompting the need for deliberate investigation in future studies. This may provide insight into the observed setting dependencies, which are postulated to be related to the user's behavior.

This chapter demonstrates for the first time the remarkable capability and feasibility of an intelligent ultrasonic aspirator that is able to distinguish between different tissue models while performing its core element – tissue resection. This is a formidable challenge, especially in the delicate area of ultra-soft brain tissue where precision is paramount. Its performance is on par with, or even surpasses, that of a human individual, marking a substantial advance in the field of neurosurgery and the development of an intelligent system based on an existing ultrasonic aspirator.

CHAPTER 7

Conclusion

In this thesis, an intelligent component for an ultrasonic aspirator was investigated, developed, and evaluated on a variety of different data and scenarios. The goal was to develop an intelligent system on the basis of an existing ultrasonic aspirator, which is able to differentiate between different tissue types encountered during surgery on human brain tissue. The approach of tissue differentiation on ultra-soft tissue with simultaneous resection of the same, made possible by the addition of intelligent methods to an existing ultrasonic aspirator concept, represents a novel approach to neurosurgery. This novel approach marks an important advancement in the field of neurosurgery, as it provides a new method for developing intraoperative decision-making aids in the future.

The system's output is derived from the electrical signals emitted by the generator associated with the ultrasonic aspirator, which change as a consequence of the interaction between the ultrasonic aspirator and the tissue. The complete process comprises two primary modules, which are hierarchically linked. The first module is contact detection, while the second is tissue differentiation. These modules are linked in such a way that tissue differentiation can only be carried out if a valid contact has been detected previously. The modularity of this approach allows for the independent consideration and investigation of each problem in terms of feasibility. In the next two paragraphs, the initial research questions I) and II) – as outlined in chapter 1 – are addressed.

Contact Detection Feasibility The feasibility of detecting whether the ultrasonic aspirator is in contact with tissue or not was successfully demonstrated, thereby confirming research question I) positively. A combination of an ensemble of deep learning-based recurrent neural networks and a recalibration of incoming data streams was used to achieve this. Scoring an F1-score of about 0.9 on the test data, the method was able to detect contact with a high degree of accuracy – even under the difficult conditions of tissue resection and a handheld ultrasonic aspirator. With the help of recalibration of incoming data, the system was able to adapt to changing conditions and maintain high performance even after long periods of time. This approach permitted the approximation of the training data distribution in an online scenario. Specific focus was placed on the detection of the initial contact, as this is the most critical phase in the tissue differentiation process which is subsequently carried out. Here the system was able to

improve the missed detection rate of initial contacts by keeping it under 0.3 %, making it reliable for a tissue differentiation process. Given the relatively modest complexity of the network architecture and recalibration process, it is reasonable to expect that the system will perform in a timely manner without any critical delays.

Tissue Differentiation Feasibility Differentiating between different tissues in terms of their stiffness while resecting them with an ultrasonic aspirator was also successfully demonstrated, confirming that research question II) can be answered affirmatively. With a steady increase in difficulty – from CNC-controlled, non-resection data to realistic resection and handheld conditions – a differentiation of a variety of tissue models was achieved with metrics such as F1-score reaching up to 0.92. The final differentiation model employed a hierarchical classification approach that enabled the differentiation between three distinct tissue stiffnesses: *very soft*, *medium soft*, and *firm*. This was achieved through the use of classic machine learning algorithms and a feature set designed to capture a variety of time-series characteristics around the contact point. A hierarchical approach allowed for the return of the less specific class *soft* in the event of uncertainty regarding the differentiation of tissues in the softer domain *very soft* and *medium soft*. The uncertainties were assessed using ensembling approaches at each level of the hierarchy. As part of the various investigations, it was discovered that the performance was dependent on the generator setting. It was found that better tissue differentiation could be achieved if the generator setting was set higher. Furthermore, the impact of the force exerted during the resection process on the outcomes was evaluated. The findings demonstrated that there was no discernible correlation between the force and the outcomes. This suggested a certain resilience of the outcomes. In order to contextualize the results, a human performance comparison study was conducted to assess the human differentiation ability of tissue stiffness. The results demonstrated that the proposed algorithms exhibited at least on-par behavior, with some instances even demonstrating superior performance – legitimizing the approach for further investigation.

Limitations and Future Work Although the viability of the tissue differentiation and contact detection modules has been demonstrated, there are still some key limitations that require attention. The current configuration of the differentiation of tissues is designed to facilitate a change in contact state, specifically from *no contact* to *contact*. Consequently, a potential change in stiffness while the ultrasonic aspirator is in contact with the tissue is not considered in the current setup. Accordingly, the current approach should be regarded as a reference system that provides information for the tissue currently under investigation at initial contact. To circumvent these limitations, heterogeneous tissue models can be employed in an initial step, exhibiting variations in transition gradients. These variations would represent clear edges, as seen in be-

nign tumors, and subtle transitions, as observed in infiltrative malignant brain tumors. Accordingly, this would be accompanied by changes to the feature generation and classification processes.

While the results are encouraging, the definitive transfer of the results obtained from tissue models to real clinical data is limited. The primary limitation is the lack of a generator type that is currently used in everyday clinical practice, which is further aggravated by the limited amount of data available. Furthermore, a lack of acquisition of healthy tissue data makes it difficult to draw any conclusions about the potential capability of the system to differentiate between healthy and tumorous tissue. Even then, it remains unclear whether the stiffness of the tissue, defined as the Young's modulus, is the sole mechanical property that should be considered for tissue differentiation or whether other properties should be taken into account. Thus, making an increase in the number of data points with an adequate generator type and the acquisition of a variety of tissue types – including healthy tissue – a priority for future work. Additionally, other mechanical properties should be considered to achieve a more comprehensive understanding between tissue properties and the ultrasonic aspirator's behavior. To ensure a system's robustness and adaptability in a clinical environment, it is important to consider external influences such as instrument angle or speed of movement. These factors were not covered in the current results but have the potential to influence outcomes. Addressing these variables – and identifying others – is essential to confirm that the system can be effectively used in various scenarios.

Moreover, the system's reliance on generator settings for tissue differentiation is not yet fully understood. While higher generator settings have shown benefits for tissue differentiation, they may not be feasible in a clinical environment due to the potential for tissue damage, especially with softer tissues, caused by increased resection strength. This raises the question of whether the system can be adapted to work with lower generator settings, which would be more suitable for clinical use, or if a specific measurement mode for tissue differentiation could be developed. This should be addressed in designated studies to unravel the intricacies of each setting's impact on the ultrasonic aspirator's behavior.

Furthermore, a hardware perspective must not be neglected. Since the ultrasonic aspirator was not built and developed as a tactile sensor, it should be evaluated which electrical components would support and improve such a system and which control technology changes could be made. The same could already be observed through hardware modifications to the cover of the ultrasonic aspirator, which eliminated irrigation effects.

Outlook This thesis primarily focused on the feasibility and individual modules of the intelligent system for contact detection and tissue differentiation. Consequently, the integrated system as a whole was not evaluated within this work. However, a demon-

strator was built as part of an accompanying project – giving a glimpse of the potential of the system. This demonstrator was presented to physicians and researchers in the field of neurosurgery, who provided positive feedback, recognizing it as a promising approach for future research and development. The demonstrator effectively showcased real-time detection of contacts and tissue differentiation in a unified system on tissue models in a handheld scenario.

Moreover, the demonstrator illustrated the potential subsequent applications that could be derived from each module of the system. For instance, the demonstrator combined the functions of the generator with the contact detection module. This combination effectively prevents unnecessary energy consumption, fluid usage, and accidental damages to surrounding objects or staff in the surgical field. It also demonstrated the capability to map the detected tissue stiffness to specific generator settings, resulting in a more automated, efficient, and intelligent system. These advancements could reduce the cognitive load on the surgeons and enhance the entire surgical workflow. Ultimately, these improvements would make the ultrasonic aspirator more user-friendly by eliminating the need for manual adjustments of the generator settings. Overall, the achievements of this thesis establish a solid and innovative foundation in the novel and cutting-edge field of tissue differentiation with an intelligent ultrasonic aspirator, paving the way for future advancements in surgical precision and technology.

Bibliography

1. Abramov I, Park MT, Belykh E, et al. Intraoperative confocal laser endomicroscopy: prospective in vivo feasibility study of a clinical-grade system for brain tumors. Journal of Neurosurgery 2022; 138(3):587–97. DOI: 10.3171/2022.5.JNS2282
2. Acerbi F, Broggi M, Schebesch KM, et al. Fluorescein-Guided Surgery for Resection of High-Grade Gliomas: A Multicentric Prospective Phase II Study (FLUOGLIO). Clinical Cancer Research 2018; 24(1):52–61. DOI: 10.1158/1078-0432.CCR-17-1184
3. Adams RP and MacKay DJC. Bayesian Online Changepoint Detection. 2007. arXiv: 0710.3742
4. Agar NY, Golby AJ, Ligon KL, et al. Development of stereotactic mass spectrometry for brain tumor surgery. Neurosurgery 2011; 68(2):280–90. DOI: 10.1227/NEU.0B013E3181FF9CBB
5. Aminikhanghahi S and Cook DJ. A survey of methods for time series change point detection. Knowledge and Information Systems 2017; 51(2):339–67. DOI: 10.1007/S10115-016-0987-Z
6. Aslani S, Dayan M, Storelli L, et al. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage 2019; 196:1–15. DOI: 10.1016/J.NEUROIMAGE.2019.03.068
7. Bai S, Kolter JZ, and Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. 2018. arXiv: 1803.01271
8. Behnan J, Stangeland B, Langella T, Finocchiaro G, Murrell W, and Brinchmann JE. Ultrasonic Surgical Aspirate is a Reliable Source For Culturing Glioblastoma Stem Cells. Scientific Reports 2016; 6:32788. DOI: 10.1038/srep32788
9. Bilston LE. Brain Tissue Mechanical Properties. Vol. 3. Springer Berlin, Heidelberg, 2010 :11–24. ISBN: 978-3-642-13890-4. DOI: 10.1007/8415_2010_36
10. Bockelmann N, Kahrs B, Kesslau D, et al. Ultrasonic Aspirator for Tissue Contact Detection: An Online Classification on Time-Series. 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2023 :1–7. DOI: 10.1109/EMBC40787.2023.10339983

11. Bockelmann N, Kesslau D, Bonsanto MM, Buschschlüter S, and Ernst F. Towards machine learning-based tissue differentiation using an ultrasonic aspirator. CARS 2021: Computer Assisted Radiology and Surgery Proceedings of the 35th International Congress and Exhibition 2021 :107–8. DOI: 10.1007/s11548-021-02375-4

12. Bockelmann N, Schetelig D, Bonsanto MM, Steffen B, and Floris E. Intelligent ultrasonic-aspirator for CNS/ tumor tissue differentiation – a feasibility study using machine learning. 73. Jahrestagung der Deutschen Gesellschaft für Neurochirurgie (DGNC) Joint Meeting mit der Griechischen Gesellschaft für Neurochirurgie 2022. DOI: 10.3205/22DGNC188

13. Bockelmann N, Schetelig D, Kesslau D, Buschschlüter S, Ernst F, and Bonsanto MM. Toward intraoperative tissue classification: exploiting signal feedback from an ultrasonic aspirator for brain tissue differentiation. International Journal of Computer Assisted Radiology and Surgery 2022; 17(9):1591–9. DOI: 10.1007/S11548-022-02713-0

14. Bösel R. Das Gehirn: ein Lehrbuch der funktionellen Anatomie für die Psychologie. W. Kohlhammer GmbH, 2006. ISBN: 9783170191839

15. Bourlard H and Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics 1988; 59(4-5):291–4. DOI: 10.1007/BF00332918

16. Bratchenko IA, Bratchenko LA, Moryatov AA, et al. In vivo diagnosis of skin cancer with a portable Raman spectroscopic device. Experimental Dermatology 2021; 30(5):652–63. DOI: 10.1111/EXD.14301

17. Breiman L. Random forests. Machine Learning 2001; 45:5–32. DOI: 10.1023/A:1010933404324

18. Brock M, Ingwersen I, and Roggendorf W. Ultrasonic aspiration in neurosurgery. Neurosurgical Review 1984; 7(2). DOI: 10.1007/BF01780701

19. Broughton D, Welling AL, Monroe EH, Pirozzi K, Schulte JB, and Clymer JW. Tissue effects in vessel sealing and transection from an ultrasonic device with more intelligent control of energy delivery. Medical Devices: Evidence and Research 2013; 6:151–4. DOI: 10.2147/MDER.S51663

20. Budday S, Sommer G, Birkl C, et al. Mechanical characterization of human brain tissue. Acta Biomaterialia 2017; 48:319–40. DOI: 10.1016/j.actbio.2016.10.036

21. Budday S, Sommer G, Haybaeck J, Steinmann P, Holzapfel GA, and Kuhl E. Rheological characterization of human brain tissue. Acta Biomaterialia 2017; 60:315–29. DOI: 10.1016/j.actbio.2017.06.024

22. Budday S, Sommer G, Holzapfel GA, Steinmann P, and Kuhl E. Viscoelastic parameter identification of human brain tissue. Journal of the Mechanical Behavior of Biomedical Materials 2017; 74:463–76. DOI: 10.1016/j.jmbbm.2017.07.014

23. Budday S, Sommer G, Paulsen F, Holzapfel GA, Steinmann P, and Kuhl E. Region- and loading-specific finite viscoelasticity of human brain tissue. Proceedings in Applied Mathematics and Mechanics 2018; 18(1):e201800169. DOI: 10.1002/PAMM.201800169

24. Budday S, Nay R, Rooij R de, et al. Mechanical properties of gray and white matter brain tissue by indentation. Journal of the Mechanical Behavior of Biomedical Materials 2015; 46:318–30. DOI: 10.1016/J.JMBBM.2015.02.024

25. Budday S, Ovaert TC, Holzapfel GA, Steinmann P, and Kuhl E. Fifty Shades of Brain: A Review on the Mechanical Testing and Modeling of Brain Tissue. Archives of Computational Methods in Engineering 2020; 27:1187–230. DOI: 10.1007/s11831-019-09352-w

26. Bunevicius A, Schregel K, Sinkus R, Golby A, and Patz S. REVIEW: MR elastography of brain tumors. NeuroImage: Clinical 2020; 25:102109. DOI: 10.1016/j.nicl.2019.102109

27. Canabal-Carbia M, Eeckhout AV, Rodriguez C, et al. Depolarizing metrics in the biomedical field: Vision enhancement and classification of biological tissues. Journal of Innovative Optical Health Sciences 2023; 16(5). DOI: 10.1142/S1793545823300045

28. Ceci M and Malerba D. Classifying web documents in a hierarchy of categories: a comprehensive study. Journal of Intelligent Information Systems 2007; 28:37–78. DOI: 10.1007/s10844-006-0003-2

29. Cepeda S, Barrena C, Arrese I, Fernandez-Pérez G, and Sarabia R. Intraoperative Ultrasonographic Elastography: A Semi-Quantitative Analysis of Brain Tumor Elasticity Patterns and Peritumoral Region. World Neurosurgery 2020; 135:e258–e270. DOI: 10.1016/j.wneu.2019.11.133

30. Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016 :785–94. DOI: 10.1145/2939672.2939785

31. Chen W, Khamis H, Birznieks I, Lepora NF, and Redmond SJ. Tactile Sensors for Friction Estimation and Incipient Slip Detection - Toward Dexterous Robotic Manipulation: A Review. IEEE Sensors Journal 2018; 18(22):9049–64. DOI: 10.1109/JSEN.2018.2868340

32. Chen Y, Gao Z, Zhang F, Wen Z, and Sun X. Recent progress in self-powered multifunctional e-skin for advanced applications. Exploration 2022; 2(1):20210112. DOI: 10.1002/EXP.20210112

33. Cho K, Merriënboer BV, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014 :1724–34. DOI: 10.3115/v1/d14-1179

34. Christ M, Braun N, Neuffer J, and Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 2018; 307:72–7. DOI: 10.1016/j.neucom.2018.03.067

35. Cimino WW and Bond LJ. Physics of ultrasonic surgery using tissue fragmentation: Part I. Ultrasound in Medicine & Biology 1996; 22(1):89–100. DOI: 10.1016/0301-5629(95)02021-7

36. Cloyd JM, Malhotra NR, Weng L, Chen W, Mauck RL, and Elliott DM. Material properties in unconfined compression of human nucleus pulposus, injectable hyaluronic acid-based hydrogels and tissue engineering scaffolds. European Spine Journal 2007; 16(11):1892–8. DOI: 10.1007/S00586-007-0443-6/FIGURES/3

37. Colen RR, Kekhia H, and Jolesz FA. Multimodality intraoperative MRI for brain tumor surgery. Expert Review of Neurotherapeutics 2010; 10(10):1545–58. DOI: 10.1586/ern.10.145

38. Cortes C, Vapnik V, and Saitta L. Support-Vector Networks. Machine Leaming 1995; 20:273–97

39. Courtecuisse H, Allard J, Kerfriden P, Bordas SP, Cotin S, and Duriez C. Real-time simulation of contact and cutting of heterogeneous soft-tissues. Medical Image Analysis 2014; 18(2):394–410. DOI: 10.1016/J.MEDIA.2013.11.001

40. Cover TM and Hart PE. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory 1967; 13(1):21–7. DOI: 10.1109/TIT.1967.1053964

41. Cox DR. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology 1958; 20(2):215–32. DOI: 10.1111/J.2517-6161.1958.TB00292.X

42. Dai W, Dai C, Qu S, Li J, and Das S. Very deep convolutional neural networks for raw waveforms. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 :421–5. DOI: 10.1109/ICASSP.2017.7952190

43. Dai Y and Armand M. Vibration-based drilling depth estimation of bone. The International Journal of Medical Robotics and Computer Assisted Surgery 2021; 17(3):e2233. DOI: 10.1002/RCS.2233

44. Demi B, Ortmaier T, and Seibold U. The touch and feel in minimally invasive surgery. IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications 2005; 2005:33–8. DOI: 10.1109/HAVE.2005.1545648

45. Dommelen JA van, Sande TP van der, Hrapko M, and Peters GW. Mechanical properties of brain tissue by indentation: Interregional variation. Journal of the Mechanical Behavior of Biomedical Materials 2010; 3(2):158–66. DOI: 10.1016/J.JMBBM.2009.09.001

46. Donnelly BR and Medige J. Shear Properties of Human Brain Tissue. Journal of Biomechanical Engineering 1997; 119(4):423–32. DOI: 10.1115/1.2798289

47. Duda RO, Hart PE, and Stork DG. Pattern Classification and Scene Analysis. Vol. 1. Wiley, 1973. ISBN: 0471223611

48. Eeckhout AV, Garcia-Caurel E, Ossikovski R, et al. Depolarization metric spaces for biological tissues classification. Journal of Biophotonics 2020; 13(8):e202000083. DOI: 10.1002/JBIO.202000083

49. Eichler AF, Chung E, Kodack DP, Loeffler JS, Fukumura D, and Jain RK. The biology of brain metastases—translation to new therapies. Nature Reviews Clinical Oncology 2011; 8(6):344–56. DOI: 10.1038/nrclinonc.2011.58

50. Eisner R, Poulin B, Szafron D, Lu P, and Greiner R. Improving protein function prediction using the hierarchical structure of the gene ontology. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2005; 2005:1–10. DOI: 10.1109/CIBCB.2005.1594940

51. Erben N, Schetelig D, Buggisch J, Bonsanto MM, Buschschlüter S, and Ernst F. Intelligent Ultrasonic Aspirator: Advancing Tissue Differentiation Through Hierarchical Classification During Hand-Held Resection. Biomedical Engineering Advances 2024; 8:100133. DOI: 10.1016/j.bea.2024.100133

52. Estes M and MacElhaney JH. Response of brain tissue to compressive loading. American Society of Mechanical Engineers Biomechanical and Human Factors Conference 1970

53. Fagni T and Sebastiani F. On the Selection of Negative Examples for Hierarchical Text Categorization. Proceedings of the 3rd Language Technology Conference 2007 :24–8

54. Fallenstein GT, Hulce VD, and Melvin JW. Dynamic mechanical properties of human brain tissue. Journal of Biomechanics 1969; 2(3):217–26. DOI: 10.1016/0021-9290(69)90079-7

55. Foersch S, Heimann A, Ayyad A, et al. Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors in vivo. PLOS ONE 2012; 7(7):e41760. DOI: 10.1371/JOURNAL.PONE.0041760

56. Franceschini G, Bigoni D, Regitnig P, and Holzapfel GA. Brain tissue deforms similarly to filled elastomers and follows consolidation theory. Journal of the Mechanics and Physics of Solids 2006; 54(12):2592–620. DOI: 10.1016/J.JMPS.2006.05.004

57. Freitas A and Carvalho A. A tutorial on hierarchical classification with applications in bioinformatics. Idea Group Publishing, 2006 :175–208. ISBN: 9781599042718. DOI: 10.4018/978-1-59904-271-8.CH007

58. Freund Y and Schapire RE. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence 1999; 14(5):771–80

59. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 2001; 29(5):1189–232. DOI: 10.1214/aos/1013203451

60. Fukunaga K. Statistical Pattern Recognition. World Scientific Publishing Company, 1993 :33–60. ISBN: 9810222769. DOI: 10.1142/9789814343138_0002

61. Galford JE and McElhaney JH. A viscoelastic study of scalp, brain, and dura. Journal of Biomechanics 1970; 3(2):211–21. DOI: 10.1016/0021-9290(70)90007-2

62. Gefen A, Gefen N, Zhu Q, Raghupathi R, and Margulies SS. Age-Dependent Changes in Material Properties of the Brain and Braincase of the Rat. Journal of Neurotrauma 2004; 20(11):1163–77. DOI: 10.1089/089771503770802853

63. Gefen A and Margulies SS. Are in vivo and in situ brain tissues mechanically similar? Journal of Biomechanics 2004; 37(9):1339–52. DOI: 10.1016/J.JBIOMECH.2003.12.032

64. Geman S, Bienenstock E, and Doursat R. Neural Networks and the Bias/Variance Dilemma. Neural Computation 1992; 4(1):1–58. DOI: 10.1162/NECO.1992.4.1.1

65. Gerard IJ, Kersten-Oertel M, Petrecca K, Sirhan D, Hall JA, and Collins DL. Brain shift in neuronavigation of brain tumors: A review. Medical Image Analysis 2017; 35:403–20. DOI: 10.1016/J.MEDIA.2016.08.007

66. Goodfellow I, Bengio Y, and Courville A. Deep Learning. MIT Press, 2016. Available from: http://www.deeplearningbook.org

67. Gutierrez-Giles A, Padilla-Castañeda MA, Alvarez-Icaza L, and Gutierrez-Herrera E. Force-Sensorless Identification and Classification of Tissue Biomechanical Parameters for Robot-Assisted Palpation. Sensors 2022; 22(22):8670. DOI: 10.3390/s22228670

68. Haar GT. Therapeutic ultrasound. European Journal of Ultrasound 1999; 9(1):3–9. DOI: 10.1016/S0929-8266(99)00013-0

69. Haddadin S, Albu-Schäffer A, Luca AD, and Hirzinger G. Evaluation of Collision Detection and Reaction for a Human-Friendly Robot on Biological Tissues. IARP International Workshop on Technical Challenges and for Dependable Robots in Human Environments 2008

70. Han Y, Liu F, and Yip MC. A 2D Surgical Simulation Framework for Tool-Tissue Interaction. 2021. arXiv: 2010.13936

71. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine 2019; 25:65–9. DOI: 10.1038/s41591-018-0268-3

72. He K, Zhang X, Ren S, and Sun J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 :770–8. DOI: 10.1109/CVPR.2016.90

73. Henzi S, Krayenbühl N, Bozinov O, Regli L, and Stienen MN. Ultrasonic aspiration in neurosurgery: comparative analysis of complications and outcome for three commonly used models. Acta Neurochirurgica 2019; 161(10). DOI: 10.1007/s00701-019-04021-0

74. Herculano-Houzel S. The human brain in numbers: A linearly scaled-up primate brain. Frontiers in Human Neuroscience 2009; 3. DOI: 10.3389/NEURO.09.031.2009

75. Hochreiter S and Schmidhuber J. Long Short-Term Memory. Neural Computation 1997; 9(8):1735–80. DOI: 10.1162/NECO.1997.9.8.1735

76. Hollon TC, Pandian B, Adapa AR, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nature Medicine 2020; 26(1):52–8. DOI: 10.1038/s41591-019-0715-9

77. Howe RD. Tactile Sensing and Control of Robotic Manipulation. Journal of Advanced Robotics 1994; 8(3):245–61

78. Ilunga-Mbuyamba E, Lindner D, Avina-Cervantes J, et al. Fusion of Intraoperative 3D B-mode and Contrast-Enhanced Ultrasound Data for Automatic Identification of Residual Brain Tumors. Applied Sciences 2017; 7(4):415. DOI: 10.3390/app7040415

79. Ioffe S and Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning 2015 :448–56

80. Jamazad A, Sedghi A, Santilli AML, et al. Improved Resection Margins in Surgical Oncology Using Intraoperative Mass Spectrometry. Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020 :44–53. DOI: 10.1007/978-3-030-59716-0_5

81. Jamin Y, Boult JK, Li J, et al. Exploring the biomechanical properties of brain malignancies and their pathologic determinants in vivo with magnetic resonance elastography. Cancer Research 2015; 75(7):1216–24. DOI: 10.1158/0008-5472.CAN-14-1997

82. Janssen NNY, Kaufmann M, Santilli A, et al. Navigated tissue characterization during skin cancer surgery. International Journal of Computer Assisted Radiology and Surgery 2020; 15. DOI: 10.1007/s11548-020-02200-4

83. Jayasudha K and Kabadi MG. Soft tissues deformation and removal simulation modelling for virtual surgery. International Journal of Intelligence and Sustainable Computing 2020; 1(1):83–100. DOI: 10.1504/IJISC.2020.104830

84. Jin X, Zhu F, Mao H, Shen M, and Yang KH. A comprehensive experimental study on material properties of human brain tissue. Journal of Biomechanics 2013; 46(16):2795–801. DOI: 10.1016/J.JBIOMECH.2013.09.001

85. Johannsmann D, Langhoff A, Bode B, et al. Towards in vivo differentiation of brain tumor versus normal tissue by means of torsional resonators. Sensors and Actuators A: Physical 2013; 190:25–31. DOI: 10.1016/j.sna.2012.10.039

86. Johnsen SF, Taylor ZA, Han L, et al. Detection and modelling of contacts in explicit finite-element simulation of soft tissue biomechanics. International Journal of Computer Assisted Radiology and Surgery 2015; 10(11):1873–91. DOI: 10.1007/S11548-014-1142-5

87. Kandel ER, Schwartz JH, Jessell TM, Siegelbaum S, Hudspeth AJ, and Mack S. Principles of Neural Science. 6th ed. McGraw Hill, 2021. ISBN: 1259642232

88. Kaneko S and Kaneko S. Fluorescence-Guided Resection of Malignant Glioma with 5-ALA. International Journal of Biomedical Imaging 2016; 2016(1):6135293. DOI: 10.1155/2016/6135293

89. Kappassov Z, Corrales JA, and Perdereau V. Tactile sensing in dexterous robot hands — Review. Robotics and Autonomous Systems 2015; 74:195–220. DOI: 10.1016/J.ROBOT.2015.07.015

90. Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 2017; 30:3149–57

91. Kent B and Rossa C. Electric impedance spectroscopy feature extraction for tissue classification with electrode embedded surgical needles through a modified forward stepwise method. Computers in Biology and Medicine 2021; 135:104522. DOI: 10.1016/J.COMPBIOMED.2021.104522

92. Kesslau D. Untersuchung der Rückkopplung von Gewebe unterschiedlicher Elastizität auf ein Ultraschallsignal. 2019

93. Kidziński Ł, Delp S, and Schwartz M. Automatic real-time gait event detection in children using deep neural networks. PLOS ONE 2019; 14(1):e0211466. DOI: 10.1371/JOURNAL.PONE.0211466

94. Kiesel B, Freund J, Reichert D, et al. 5-ALA in Suspected Low-Grade Gliomas: Current Role, Limitations, and New Approaches. Frontiers in Oncology 2021; 11:699301. DOI: 10.3389/fonc.2021.699301

95. Kingma DP and Ba J. Adam: A Method for Stochastic Optimization. 2017. arXiv: 1412.6980

96. Kiritchenko S, Matwin S, and Famili F. Functional Annotation of Genes Using Hierarchical Text Categorization. Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics 2005

97. Kleesattel C and Gladwell GM. The contact - impedance meter-1. Ultrasonics 1968; 6(3):175–80. DOI: 10.1016/0041-624X(68)90280-1

98. Koch C. Thermische Wirkungen von Ultraschall. Ultraschall in der Medizin 2001; 22(3):146–52. DOI: 10.1055/S-2001-15245

99. Koller D and Sahami M. Hierarchically classifying documents using very few words. Proceedings of the Fourteenth International Conference on Machine Learning 1997 :170–8

100. Krizhevsky A, Sutskever I, and Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 2012; 25

101. Kruse SA, Rose GH, Glaser KJ, et al. Magnetic resonance elastography of the brain. NeuroImage 2008; 39(1):231–7. DOI: 10.1016/J.NEUROIMAGE.2007.08.030

102. Kühnapfel U, Çakmak HK, and Maaß H. Endoscopic surgery training using virtual reality and deformable tissue simulation. Computers & Graphics 2000; 24(5):671–82. DOI: 10.1016/S0097-8493(00)00070-4

103. Kyriacou SK, Mohamed A, Miller K, and Neff S. Brain Mechanics for Neurosurgery: Modeling Issues. Biomechanics and Modeling in Mechanobiology 2002; 1(2):151–64. DOI: 10.1007/S10237-002-0013-0

104. Latt WT, Newton RC, Visentini-Scarzanella M, et al. A Hand-held Instrument to Maintain Steady Tissue Contact during Probe-Based Confocal Laser Endomicroscopy. IEEE Transactions on Biomedical Engineering 2011; 58(9). DOI: 10.1109/TBME.2011.2162064

105. LeCun Y, Boser B, Denker JS, et al. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1989; 1(4):541–51. DOI: 10.1162/NECO.1989.1.4.541

106. Lindahl OA, Constantinou CE, Eklund A, Murayama Y, Hallberg P, and Omata S. Tactile resonance sensors in medicine. Journal of Medical Engineering and Technology 2009; 33(4):263–73. DOI: 10.1080/03091900802491188

107. Lindahl OA, Backlund T, Ramser K, Liv P, Ljungberg B, and Bergh A. A tactile resonance sensor for prostate cancer detection - Evaluation on human prostate tissue. Biomedical Physics and Engineering Express 2021; 7(2). DOI: 10.1088/2057-1976/abe681

108. Liu L, Jiang H, He P, et al. On the Variance of the Adaptive Learning Rate and Beyond. Proceedings of the 8th International Conference on Learning Representations 2020

109. Louis DN, Ohgaki H, Wiestler OD, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathologica 2007; 114(2):97–109. DOI: 10.1007/S00401-007-0243-4

110. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathologica 2016; 131(6):803–20. DOI: 10.1007/S00401-016-1545-1

111. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology 2021; 23(8):1231–51. DOI: 10.1093/NEUONC/NOAB106

112. Machado MMP, Voda A, Besançon G, Becq G, Kahane P, and David O. Brain tissue classification from stereoelectroencephalographic recordings. Journal of Neuroscience Methods 2022; 365:109375. DOI: 10.1016/J.JNEUMETH.2021.109375

113. MacManus DB, Pierrat B, Murphy JG, and Gilchrist MD. Dynamic mechanical properties of murine brain tissue using micro-indentation. Journal of Biomechanics 2015; 48(12):3213–8. DOI: 10.1016/J.JBIOMECH.2015.06.028

114. McFaline-Figueroa JR and Lee EQ. Brain Tumors. The American Journal of Medicine 2018; 131(8):874–82. DOI: 10.1016/J.AMJMED.2017.12.039

115. McNeill KA. Epidemiology of Brain Tumors. Neurologic Clinics 2016; 34(4):981–98. DOI: 10.1016/J.NCL.2016.06.014

116. Milsom JW, Trencheva K, Sonoda T, Nandakumar G, Shukla PJ, and Lee S. A prospective trial evaluating the clinical performance of a novel surgical energy device in laparoscopic colon surgery. Surgical Endoscopy 2015; 29:1161–6. DOI: 10.1007/s00464-014-3783-4

117. Min S, Lidde J, Raue N, and Dornfeld D. Acoustic emission based tool contact detection for ultra-precision machining. CIRP Annals 2011; 60(1):141–4. DOI: 10.1016/J.CIRP.2011.03.079

118. Mohammed S, Dinesan M, and Ajayakumar T. Survival and quality of life analysis in glioblastoma multiforme with adjuvant chemoradiotherapy: a retrospective study. Reports of Practical Oncology and Radiotherapy 2022; 27(6):1026. DOI: 10.5603/RPOR.A2022.0113

119. Müller W, Fritzsch G, Füssel J, and Aurich H. Ultrasound instruments for surgery. Technical aspects and selected uses. Ultraschall in der Medizin 1990; 11(2):76–80. DOI: 10.1055/S-2007-1011534/BIB

120. Nair V and Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on Machine Learning 2010 :807–14

121. Navas-Palencia G. Optimal binning: mathematical programming formulation. 2022. arXiv: 2001.08025

122. Nusrat I and Jang SB. A Comparison of Regularization Techniques in Deep Neural Networks. Symmetry 2018; 10(11):648. DOI: 10.3390/SYM10110648

123. O'Daly BJ, Morris E, Gavin GP, O'Byrne JM, and McGuinness GB. High-power low-frequency ultrasound: A review of tissue dissection and ablation in medicine and surgery. Journal of Materials Processing Technology 2008; 200(1-3):38–58. DOI: 10.1016/J.JMATPROTEC.2007.11.041

124. Omata S, Murayama Y, and Constantinou CE. Real time robotic tactile sensor system for the determination of the physical properties of biomaterials. Sensors and Actuators A: Physical 2004; 112(2-3):278–85. DOI: 10.1016/J.SNA.2004.01.038

125. Omata S and Terunuma Y. New tactile sensor like the human hand and its applications. Sensors and Actuators A: Physical 1992; 35:9–15. DOI: 10.1016/0924-4247(92)87002-X

126. Ostler D, Seibold M, Fuchtmann J, et al. Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions. International Journal of Computer Assisted Radiology and Surgery 2020; 15(5):771–9. DOI: 10.1007/s11548-020-02146-7

127. Packer M, Fishkind WJ, Fine IH, Seibel BS, and Hoffman RS. The physics of phaco: A review. Journal of Cataract and Refractive Surgery 2005; 31(2):424–31. DOI: 10.1016/J.JCRS.2004.11.027

128. Pagoli A, Chapelle F, Ramon JAC, Mezouar Y, and Lapusta Y. Review of soft fluidic actuators: classification and materials modeling analysis. Smart Materials and Structures 2021; 31(1):013001. DOI: 10.1088/1361-665X/ac383aï

129. Palmieri G, Cofano F, Salvati LF, et al. Fluorescence-Guided Surgery for High-Grade Gliomas: State of the Art and New Perspectives. Technology in Cancer Research and Treatment 2021; 20. DOI: 10.1177/15330338211021605

130. Park BJ, Kim HK, Sade B, and Lee JH. Epidemiology. Meningiomas 2009 :11–4. DOI: 10.1007/978-1-84628-784-8_2

131. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30

132. Prange MT and Margulies SS. Regional, Directional, and Age-Dependent Properties of the Brain Undergoing Large Deformation. Journal of Biomechanical Engineering 2002; 124(2):244–52. DOI: 10.1115/1.1449907

133. Prechelt L. Automatic early stopping using cross validation: quantifying the criteria. Neural Networks 1998; 11(4):761–7. DOI: 10.1016/S0893-6080(98)00010-0

134. Psota E, Carlson J, Armijo PR, et al. End-Effector Contact and Force Detection for Miniature Autonomous Robots Performing Lunar and Expeditionary Surgery. Military Medicine 2021; 186:281–7. DOI: 10.1093/MILMED/USAA443

135. Puangmali P, Althoefer K, Seneviratne LD, Murphy D, and Dasgupta P. State-of-the-art in force and tactile sensing for minimally invasive surgery. IEEE Sensors Journal 2008; 8(4):371–80. DOI: 10.1109/JSEN.2008.917481

136. Rashid B, Destrade M, and Gilchrist MD. Mechanical characterization of brain tissue in compression at dynamic strain rates. Journal of the Mechanical Behavior of Biomedical Materials 2012; 10:23–38. DOI: 10.1016/J.JMBBM.2012.01.022

137. Rashid B, Destrade M, and Gilchrist MD. Temperature effects on brain tissue in compression. Journal of the Mechanical Behavior of Biomedical Materials 2012; 14:113–8. DOI: 10.1016/J.JMBBM.2012.04.005

138. Rasmussen CE and Williams CKI. Gaussian Processes for Machine Learning. MIT Press, 2006. ISBN: 026218253X

139. Reiss-Zimmermann M, Streitberger KJ, Sack I, et al. High Resolution Imaging of Viscoelastic Properties of Intracranial Tumours by Multi-Frequency Magnetic Resonance Elastography. Clinical Neuroradiology 2015; 25:371–8. DOI: 10.1007/s00062-014-0311-9

140. Rengaraju V, Silva A da, Sack I, and Kargel C. A basic study of ultrasonic shear wave elastography in tissue-mimicking phantoms. IEEE International Workshop on Medical Measurements and Applications 2009 :55–9. DOI: 10.1109/MEMEA.2009.5167954

141. Restelli F, Mathis AM, Höhne J, et al. Confocal laser imaging in neurosurgery: A comprehensive review of sodium fluorescein-based CONVIVO preclinical and clinical applications. Frontiers in Oncology 2022; 12:998384. DOI: 10.3389/FONC.2022.998384

142. Restelli F, Mazzapicchi E, Pollo B, et al. A new study protocol for in-vivo assessment of tumor diagnosis and microscopic tumor infiltration at the resection cavity in central nervous system tumors by a new miniature confocal endomicroscope (CONVIVO system). Journal of Neurosurgical Sciences 2023; 67(3):280–7. DOI: 10.23736/S0390-5616.22.05906-9

143. Reyns N, Leroy HA, Delmaire C, Derre B, Le-Rhun E, and Lejeune JP. Intraoperative MRI for the management of brain lesions adjacent to eloquent areas. Neurochirurgie 2017; 63(3):181–8. DOI: 10.1016/j.neuchi.2016.12.006

144. Ritschel K, Pechlivanis I, and Winter S. Brain tumor classification on intraoperative contrast-enhanced ultrasound. International Journal of Computer Assisted Radiology and Surgery 2015; 10(5):531–40. DOI: 10.1007/s11548-014-1089-6

145. Ronneberger O, Fischer P, and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI) 2015; 9351:234–41

146. Rosenblatt F. The Perceptron — A Perceiving and Recognizing Automaton. Cornell Aeronautical Laborator, 1957

147. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 1958; 65(6):386–408. DOI: 10.1037/H0042519

148. Ruderman S, Mueller S, Gomes A, Rogers J, and Backman V. Method of detecting tissue contact for fiber-optic probes to automate data acquisition without hardware modification. Biomedical Optics Express 2013; 4(8):1401. DOI: 10.1364/BOE.4.001401

149. Rumelhart DE, Hinton GE, and Williams RJ. Learning representations by back-propagating errors. Nature 1986; 323:533–6. DOI: 10.1038/323533a0

150. Sanchez J, Corrales JA, Bouzgarrou BC, and Mezouar Y. Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. The International Journal of Robotics Research 2018; 37(7):688–716. DOI: 10.1177/0278364918779698

151. Sandykbayeva D, Kappassov Z, and Orazbayev B. VibroTouch: Active Tactile Sensor for Contact Detection and Force Sensing via Vibrations. Sensors 2022; 22(17):6456. DOI: 10.3390/S22176456/S1

152. Santilli AM, Jamzad A, Janssen NN, et al. Perioperative margin detection in basal cell carcinoma using a deep learning framework: a feasibility study. International Journal of Computer Assisted Radiology and Surgery 2020; 15(5):887–96. DOI: 10.1007/s11548-020-02152-9

153. Santilli AM, Jamzad A, Sedghi A, et al. Domain adaptation and self-supervised learning for surgical margin detection. International Journal of Computer Assisted Radiology and Surgery 2021; 16(5):861–9. DOI: 10.1007/S11548-021-02381-6

154. Santilli AM, Jamzad A, Sedghi A, et al. Self-supervised learning for detection of breast cancer in surgical margins with limited data. IEEE 18th International Symposium on Biomedical Imaging (ISBI) 2021 :980–4. DOI: 10.1109/ISBI48211.2021.9433829

155. Schäfer KC, Balog J, Szaniszló T, et al. Real time analysis of brain tissue by direct combination of ultrasonic surgical aspiration and sonic spray mass spectrometry. Analytical Chemistry 2011; 83(20):7729–35. DOI: 10.1021/ac201251s

156. Secker A, Davies MN, Freitas AA, Timmis J, Mendao M, and Flower DR. An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function. Expert Update (Magazine of the British Computer Society's Specialist Group on AI) 2007; 9(3):17–22

157. Senders JT, Muskens IS, Schnoor R, et al. Agents for fluorescence-guided glioma surgery: a systematic review of preclinical and clinical results. Acta Neurochirurgica 2017; 159:151–67. DOI: 10.1007/s00701-016-3028-5

158. Shuck LZ and Advani SH. Rheological Response of Human Brain Tissue in Shear. Journal of Basic Engineering 1972; 94(4):905–11. DOI: 10.1115/1.3425588

159. Silla CN and Freitas AA. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery 2011; 22(1-2):31–72. DOI: 10.1007/s10618-010-0175-9

160. Smith JS, Chang EF, Lamborn KR, et al. Role of Extent of Resection in the Long-Term Outcome of Low-Grade Hemispheric Gliomas. Journal of Clinical Oncology 2016; 26(8):1338–45. DOI: 10.1200/JCO.2007.13.9337

161. Srivastava N, Hinton G, Krizhevsky A, and Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014; 15:1929–58

162. Strenge P, Lange B, Grill C, et al. Characterization of brain tumor tissue with 1310 nm optical coherence tomography. Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXV 2021; 11630:116301T. DOI: 10.1117/12.2578409

163. Strenge P, Lange B, Grill C, et al. Creating a depth-resolved OCT-dataset for supervised classification based on ex vivo human brain samples. Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXV 2021; 11630:116301S. DOI: 10.1117/12.2578391

164. Strenge P, Lange B, Grill C, et al. Ex vivo and in vivo imaging of human brain tissue with different OCT systems. Optical Coherence Imaging Techniques and Imaging in Scattering Media III 2019; 11078:110781C. DOI: 10.1117/12.2526932

165. Stroop R, Nakamura M, Schoukens J, and Uribe DO. Tactile sensor-based real-time clustering for tissue differentiation. International Journal of Computer Assisted Radiology and Surgery 2019; 14:129–37. DOI: 10.1007/s11548-018-1869-5

166. Stummer W, Reulen HJ, Meinel T, et al. Extent of resection and survival in glioblastoma multiforme: Identification of and adjustment for bias. Neurosurgery 2008; 62(3):564–74. DOI: 10.1227/01.NEU.0000317304.31579.17

167. Sühn T, Esmaeili N, Spiller M, et al. Vibro-acoustic sensing of tissue-instrument-interactions allows a differentiation of biological tissue in computerised palpation. Computers in Biology and Medicine 2023; 164:107272. DOI: 10.1016/J.COMPBIOMED.2023.107272

168. Sühn T, Pandey A, Friebe M, Illanes A, Boese A, and Lohman C. Acoustic sensing of tissue-tool interactions-Potential applications in arthroscopic surgery. Current Directions in Biomedical Engineering 2020; 6(3):595–8. DOI: 10.1515/CDBME-2020-3152

169. Sun A and Lim EP. Hierarchical text classification and evaluation. IEEE International Conference on Data Mining 2001 :521–8. DOI: 10.1109/ICDM.2001.989560

170. Tanaka Y, Yu Q, Doumoto K, et al. Development of a real-time tactile sensing system for brain tumor diagnosis. International Journal of Computer Assisted Radiology and Surgery 2010; 5(4):359–67. DOI: 10.1007/s11548-010-0426-7

171. Tavakoli M, Patel R, and Moallem M. Robotic suturing forces in the presence of haptic feedback and sensory substitution. Proceedings of 2005 IEEE Conference on Control Applications 2005 :1–6. DOI: 10.1109/CCA.2005.1507091

172. Tejo-Otero A, Fenollosa-Artés F, Achaerandio I, et al. Soft-Tissue-Mimicking Using Hydrogels for the Development of Phantoms. Gels 2022; 8(1):40. DOI: 10.3390/GELS8010040

173. Thai MT, Phan PT, Hoang TT, Wong S, Lovell NH, and Do TN. Advanced Intelligent Systems for Surgical Robotics. Advanced Intelligent Systems 2020; 2(8):1900138. DOI: 10.1002/AISY.201900138

174. Tkachenko M, Malyuk M, Holmanyuk A, and Liubimov N. Label Studio: Data labeling software. 2022. Available from: https://github.com/heartexlabs/label-studio

175. Turner SL, Rabani A, Axinte DA, and King CW. Dynamic ultrasonic contact detection using acoustic emissions. Ultrasonics 2014; 54(3):749–53. DOI: 10.1016/J.ULTRAS.2013.10.002

176. Uff CE, Garcia L, Fromageau J, Dorward N, and Bamber JC. Real-time ultrasound elastography in neurosurgery. IEEE International Ultrasonics Symposium 2009 :467–70. DOI: 10.1109/ULTSYM.2009.5441629

177. Vargo M. Brain tumor rehabilitation. American Journal of Physical Medicine and Rehabilitation 2011; 90(5):50–62. DOI: 10.1097/PHM.0B013E31820BE31F

178. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Advances in Neural Information Processing Systems 2017; 30

179. Veit I. Technische Akustik Grundlagen der physikalischen, gehörbezogenen Elektro- und Bauakustik. Vol. 7. Vogel Buchverlag, 1996. ISBN: 978-3-8343-3282-0

180. Weickenmeier J, Rooij R de, Budday S, Steinmann P, Ovaert TC, and Kuhl E. Brain stiffness increases with myelin content. Acta Biomaterialia 2016; 42:265–72. DOI: 10.1016/J.ACTBIO.2016.07.040

181. Whittall KP, MacKay AL, Graeb DA, Nugent RA, Li DK, and Paty DW. In vivo measurement of T2 distributions and water contents in normal human brain. Magnetic Resonance in Medicine 1997; 37(1):34–43. DOI: 10.1002/MRM.1910370107

182. Williams RW and Herrup K. The control of neuron number. Annual Review of Neuroscience 1988; 11:423–53. DOI: 10.1146/ANNUREV.NE.11.030188.002231

183. Yoshimoto S, Kuroda Y, Imura M, Oshiro O, and Sato K. Smart sensing of tool/tissue interaction by resistive coupling. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2013 :628–31. DOI: 10.1109/EMBC.2013.6609578

184. Youn SH, Shin KY, Choi A, and Mun JH. Evaluation of Robust Classifier Algorithm for Tissue Classification under Various Noise Levels. ETRI Journal 2017; 39(1):87–96. DOI: 10.4218/ETRIJ.17.0116.0113

185. Youn SH, Sim T, Choi A, et al. Multi-class biological tissue classification based on a multi-classifier: Preliminary study of an automatic output power control for ultrasonic surgical units. Computers in Biology and Medicine 2015; 61:92–100. DOI: 10.1016/J.COMPBIOMED.2015.03.021

186. Zhang Y, Ju F, Wei X, Wang D, and Wang Y. A Piezoelectric Tactile Sensor for Tissue Stiffness Detection with Arbitrary Contact Angle. Sensors 2020; 20(22):6607. DOI: 10.3390/S20226607

A Data

This appendix presents comprehensive details regarding the data utilized within this thesis.

Table A.1: Average Young's Modulus for different settings for data set D6.

Setting	Average Young's Modulus in Pa
high	639.0
high	872.0
high	1077.6
high	1242.7
high	1316.0
high	3020.0
high	3111.5
high	4890.0
high	5365.0
high	6180.0
high	6387.0
high	6606.7
high	7330.0
high	8756.0
high	10636.0
medium	639.0
medium	872.0
medium	879.3
medium	931.0
medium	1062.0
medium	1077.6
medium	1089.0
medium	1091.0
medium	1194.0
medium	1235.1
medium	1242.7
medium	1316.0
medium	3020.0
medium	3111.5

medium	4290.0
medium	5167.0
medium	5206.0
medium	5365.0
medium	6606.7
medium	8756.0
medium	10636.0
low	639.0
low	872.0
low	1077.6
low	1242.7
low	1316.0
low	3020.0
low	3111.5
low	4890.0
low	5365.0
low	6180.0
low	6387.0
low	6606.7
low	7330.0
low	8756.0
low	10636.0

B Results

This appendix presents additional tables containing detailed results from the experiments conducted throughout this thesis. Metrics used are accuracy (ACC), balanced accuracy (BA), F1-score (F1), positive predictive value (PPV), true positive rate (TPR), as well as their hierarchical counterparts (hF1, hPPV, hTPR). Specific metric for start and end areas of contacts are also included as time to contact (TT) and miss rate (MSS).

Table B.1: Results of the three methods on the test data set grouped by setting while being trained with and without setting dependence. A setting dependent result means each setting has its designated trained model instead of a single model for all settings.

Method	Setting Dependent	Setting	BA	F1	PPV	TPR	Start			End		
							TT	ACC	MSS	TT	ACC	MSS
ROCTh	False	Low	0.872	0.871	0.874	0.870	14.3	0.862	8.6	26.0	0.775	16.8
		Medium	0.886	0.888	0.890	0.887	11.8	0.875	2.1	21.3	0.789	7.6
		High	0.783	0.834	0.868	0.846	12.5	0.845	9.0	21.3	0.717	22.9
	True	Low	0.881	0.883	0.883	0.882	13.2	0.870	9.1	25.2	0.788	14.7
		Medium	0.886	0.892	0.892	0.891	11.7	0.873	2.1	20.8	0.789	7.6
		High	0.828	0.863	0.871	0.868	11.4	0.880	2.2	19.4	0.767	10.3
DynTh	False	Low	0.874	0.875	0.877	0.874	12.8	0.877	14.3	28.1	0.776	27.9
		Medium	0.921	0.924	0.924	0.924	8.8	0.925	8.0	18.8	0.858	13.9
		High	0.828	0.864	0.874	0.869	7.3	0.852	25.1	29.0	0.711	42.2
	True	Low	0.859	0.865	0.865	0.865	11.2	0.865	18.4	27.1	0.773	29.4
		Medium	0.925	0.932	0.932	0.932	8.2	0.928	8.0	19.1	0.857	13.9
		High	0.849	0.877	0.880	0.879	8.0	0.880	17.9	29.7	0.731	34.5
DynRNN	False	Low	0.935	0.935	0.935	0.935	10.9	0.934	0.0	18.3	0.858	1.0
		Medium	0.868	0.869	0.873	0.868	11.3	0.907	2.9	18.4	0.842	1.7
		High	0.819	0.846	0.848	0.849	7.1	0.926	0.0	20.6	0.744	4.5
	True	Low	0.813	0.790	0.833	0.789	17.7	0.820	14.9	20.0	0.846	3.0
		Medium	0.846	0.829	0.854	0.826	16.3	0.861	2.1	18.2	0.851	0.4
		High	0.675	0.735	0.808	0.769	10.2	0.835	4.5	24.9	0.640	31.4

Table B.2: Results of the flat and hierarchical classification approaches with enforced confidence thresholds on test data.

Method	Confidence Threshold	BA	F1	PPV	TPR	hF1	hPPV	hTPR
Flat	40	0.763	0.762	0.771	0.763	-	-	-
	50	0.763	0.762	0.771	0.763	-	-	-
	60	0.778	0.779	0.784	0.780	-	-	-
	70	0.814	0.813	0.818	0.815	-	-	-
	80	0.842	0.849	0.852	0.851	-	-	-
	90	0.846	0.850	0.854	0.853	-	-	-
	100	0.880	0.891	0.896	0.893	-	-	-
MLNP	50	0.780	0.781	0.794	0.780	0.823	0.828	0.823
	60	0.785	0.785	0.796	0.783	0.823	0.829	0.822
	70	0.804	0.804	0.813	0.802	0.833	0.840	0.830
	80	0.823	0.826	0.836	0.824	0.853	0.860	0.849
	90	0.833	0.841	0.849	0.840	0.871	0.880	0.866
	100	0.890	0.896	0.899	0.896	0.914	0.917	0.911
NMLNP	60	0.839	0.788	0.798	0.786	0.822	0.830	0.819
	70	0.853	0.817	0.825	0.815	0.825	0.847	0.809
	80	0.867	0.845	0.854	0.844	0.836	0.870	0.809
	90	0.875	0.868	0.874	0.867	0.843	0.893	0.804
	100	0.918	0.920	0.922	0.920	0.866	0.930	0.816

Table B.3: Results of the human performance comparison study with the distinction of differente participant groups and batch IDs over different metrics. Note that Non-Expert and Expert are subgroups of the participant group All, while Specialist and Assistant Physician are subgroups of the Expert group. Ts represent all test batch IDs. Ts ∩ Tr represent the test batch IDs that are also present in the training data. Ts \ Tr represent the test batch IDs that are not present in the training data.

Participant Group	Batches	ACC	F1	PPV	TPR
All	Ts	0.708	0.701	0.700	0.708
All	$\mathrm{Ts} \setminus \mathrm{Tr}$	0.663	0.668	0.753	0.663
All	$\mathrm{Ts} \cap \mathrm{Tr}$	0.764	0.765	0.822	0.764
Expert	Ts	0.735	0.730	0.728	0.735
Expert	$\mathrm{Ts} \setminus \mathrm{Tr}$	0.662	0.668	0.722	0.662
Expert	$\mathrm{Ts} \cap \mathrm{Tr}$	0.827	0.827	0.887	0.827
Non-Expert	Ts	0.692	0.684	0.683	0.692
Non-Expert	$\mathrm{Ts} \setminus \mathrm{Tr}$	0.664	0.668	0.780	0.664
Non-Expert	$\mathrm{Ts} \cap \mathrm{Tr}$	0.727	0.728	0.783	0.727
Specialist	Ts	0.762	0.760	0.758	0.762
Specialist	$\mathrm{Ts} \setminus \mathrm{Tr}$	0.686	0.696	0.727	0.686
Specialist	$\mathrm{Ts} \cap \mathrm{Tr}$	0.857	0.861	0.909	0.857
Assistant Physician	Ts	0.704	0.695	0.693	0.704
Assistant Physician	$\mathrm{Ts} \setminus \mathrm{Tr}$	0.633	0.628	0.717	0.633
Assistant Physician	$\mathrm{Ts} \cap \mathrm{Tr}$	0.792	0.787	0.864	0.792

Table B.4: The top 25 *tsfresh* features ranked on Pearson correlation between feature and force per *soft* and *firm* tissue class. Bold entries indicate the highest ranked common feature between the two classes. Details on the features can be found in the *tsfresh* documentation [34] and are omitted here for brevity.

soft		*firm*	
Feature	Correlation	Feature	Correlation
voltage number_crossing_m m_1	0.380	voltage sample_entropy	-0.524
voltage fft_coefficient attr_"abs" coeff_13	0.374	voltage change_quantiles f_agg_"mean" isabs_True qh_0.4 ql_0.2	-0.500
voltage fft_coefficient attr_"abs" coeff_59	-0.367	voltage approximate_entropy m_2 r_0.1	-0.492
voltage change_quantiles f_agg_"var" isabs_True qh_0.8 ql_0.2	0.353	voltage change_quantiles f_agg_"mean" isabs_True qh_0.4 ql_0.0	-0.49
voltage fft_coefficient attr_"abs" coeff_61	0.342	voltage change_quantiles f_agg_"mean" isabs_True qh_0.2 ql_0.0	-0.488
voltage change_quantiles f_agg_"var" isabs_True qh_0.8 ql_0.4	0.336	voltage change_quantiles f_agg_"var" isabs_False qh_0.2 ql_0.0	-0.488
frequency agg_linear_trend attr_"rvalue" chunk_len_5 f_agg_"var"	-0.328	voltage change_quantiles f_agg_"var" isabs_False qh_0.4 ql_0.0	-0.487
voltage fft_coefficient attr_"angle" coeff_80	-0.316	voltage fft_coefficient attr_"angle" coeff_80	-0.316
voltage approximate_entropy m_2 r_0.7	0.314	voltage autocorrelation lag_3	0.480
voltage change_quantiles f_agg_"var" isabs_False qh_0.8 ql_0.4	0.313	voltage approximate_entropy m_2 r_0.3	-0.471
voltage fourier_entropy bins_100	0.308	voltage change_quantiles f_agg_"mean" isabs_True qh_0.6 ql_0.0	-0.465
voltage change_quantiles f_agg_"var" isabs_True qh_1.0 ql_0.0	0.307	voltage autocorrelation lag_4	0.459
voltage agg_linear_trend attr_"stderr" chunk_len_5 f_agg_"var"	0.303	voltage friedrich_coefficients coeff_0 m_3 r_30	0.453
voltage agg_linear_trend attr_"stderr" chunk_len_10 f_agg_"var"	0.302	voltage mean_abs_change	-0.451
voltage change_quantiles f_agg_"var" isabs_True qh_1.0 ql_0.2	0.300	voltage absolute_sum_of_changes	-0.451
voltage fft_coefficient attr_"imag" coeff_14	-0.298	voltage change_quantiles f_agg_"mean" isabs_True qh_1.0 ql_0.0	-0.451
voltage change_quantiles f_agg_"var" isabs_False qh_0.8 ql_0.2	0.297	voltage change_quantiles f_agg_"var" isabs_False qh_0.6 ql_0.0	-0.451
voltage fft_coefficient attr_"abs" coeff_57	0.297	**voltage change_quantiles f_agg_"mean" isabs_True qh_0.8 ql_0.0**	-0.449
voltage agg_linear_trend attr_"intercept" chunk_len_50 f_agg_"var"	0.294	voltage approximate_entropy m_2 r_0.5	-0.448
voltage fft_coefficient attr_"abs" coeff_33	0.289	voltage change_quantiles f_agg_"var" isabs_False qh_1.0 ql_0.0	-0.441
frequency ratio_value_number_to_time_series_length	-0.276	voltage change_quantiles f_agg_"mean" isabs_True qh_1.0 ql_0.2	-0.440
voltage change_quantiles f_agg_"var" isabs_True qh_1.0 ql_0.4	0.276	voltage cid_ce normalize_False	-0.440
voltage fft_coefficient attr_"abs" coeff_11	0.276	voltage cid_ce normalize_True	-0.440
voltage change_quantiles f_agg_"mean" isabs_True qh_0.8 ql_0.0	0.273	voltage change_quantiles f_agg_"mean" isabs_True qh_1.0 ql_0.6	-0.437
voltage change_quantiles f_agg_"var" isabs_False qh_1.0 ql_0.0	0.267	voltage autocorrelation lag_5	0.433

C Algorithms and Parameters

This appendix presents hyperparameters and pseudocode for the algorithms and models referenced throughout the thesis. These detailed listings provide a clear reference for the specific settings and implementation details used in the thesis, complementing the discussions in the main chapters.

Table C.1: Hyperparameters of neural network architecture. n = number of input features, AE = Autoencoder, Clf = Classifier, FC = Fully connected

	Layer Type	Size	Non-linearity
AE - Hidden Encoder	FC	32	ReLU
AE - Latent Space	FC	9	None
AE - Hidden Decoder	FC	32	ReLU
AE - Output	FC	n	None
Clf - Hidden	FC	4	ReLU
Clf - Output	FC	1	None

Input: voltage sequence, $threshold$
Output: prediction
Initialize variables: $contact\_index \leftarrow -1$, $contact\_state \leftarrow 0$,
$dyn\_threshold \leftarrow None$;
Initialize buffers: $\mathbf{b}^{c}$, $\mathbf{b}^{nc}$, $prediction$;
Initialize flags: $contact\_detected \leftarrow False$,
$no\_contact\_buffer\_filled \leftarrow False$, $contact\_buffer\_filled \leftarrow False$;
for i *in length of voltage sequence* **do**
  if *not* $contact\_detected$ **then**
    Calculate $value$ to test against $threshold$; /* Equation 5.2 */
    if $value > threshold$ **then**
      $contact\_detected \leftarrow True$;
      $contact\_index \leftarrow i$;
      $contact\_state \leftarrow 1$;
    end
  end
  if $contact\_detected$ *and not* $no\_contact\_buffer\_filled$ *and*
  $i = contact\_index$ **then**
    Push previous voltage values to no contact buffer $\mathbf{b}^{nc}$;
    $no\_contact\_buffer\_filled \leftarrow True$;
  end
  if $contact\_detected$ *and* $i > (contact\_index + 10)$ *and*
  $i <= (contact\_index + 110)$ *and not* $contact\_buffer\_filled$ **then**
    Push current voltage value to contact buffer $\mathbf{b}^{c}$;
    if *length of contact buffer* $= 100$ **then**
      $contact\_buffer\_filled \leftarrow True$;
    end
  end
  if $contact\_buffer\_filled$ *and* $no\_contact\_buffer\_filled$ **then**
    Calculate $dyn\_threshold$; /* Equation 5.4 */
  end
  if $dyn\_threshold \neq None$ **then**
    if *current voltage value* $> dyn\_threshold$ **then**
      $contact\_state \leftarrow 1$;
    end
    else
      $contact\_state \leftarrow 0$;
    end
  end
  if $contact\_state = 0$ *and* $contact\_detected$ **then**
    Reset $\mathbf{b}^{c}$, $\mathbf{b}^{nc}$ and flags;
    $dyn\_threshold \leftarrow None$;
  end
  Push $contact\_state$ to $prediction$;
end

Algorithm 1: Pseudo-code for contact detection using *DynTh*.

Input: voltage sequence, RNN
Output: prediction
Initialize variables: $contact\_index \leftarrow -1$, $contact\_state \leftarrow 0$, $cal\_value \leftarrow 0$, $cal\_value\_prev \leftarrow 0$;
Initialize buffers: $\mathbf{b}^{c}$, $\mathbf{b}^{nc}$, $prediction$;
Initialize flags: $contact\_detected \leftarrow False$, $no\_contact\_buffer\_filled \leftarrow False$, $contact\_buffer\_filled \leftarrow False$;
for i *in length of voltage sequence* **do**
    $sequence \leftarrow voltage\_sequence[i - 100 : i]$; `/* 100 data points */`
    $sequence \leftarrow sequence - cal\_value\_prev - cal\_value$;
    Calculate $differentiated\ sequence$;
    $contact\_state \leftarrow RNN(sequence, differentiated\ sequence)$;
    if $contact\_state = 1$ *and not* $contact\_detected$ **then**
        $contact\_detected \leftarrow True$;
        $contact\_index \leftarrow i$;
    end
    if $contact\_state = 0$ *and* $contact\_detected$ **then**
        $contact\_detected \leftarrow False$;
        Reset $\mathbf{b}^{c}$, $\mathbf{b}^{nc}$ and its flags;
    end
    if $contact\_detected$ *and not* $non\_contact\_buffer\_filled$ *and* $i = contact\_index$ **then**
        Push calibrated previous voltage values to no contact buffer $\mathbf{b}^{nc}$;
        $no\_contact\_buffer\_filled \leftarrow True$;
    end
    if $contact\_detected$ *and* $i > (contact\_index + 10)$ *and* $i <= (contact\_index + 110)$ *and not* $contact\_buffer\_filled$ **then**
        Push calibrated current voltage value to contact buffer $\mathbf{b}^{c}$;
        if *length of contact buffer* $= 75$ **then**
            $contact\_buffer\_filled \leftarrow True$;
        end
    end
    if $contact\_buffer\_filled$ *and* $non\_contact\_buffer\_filled$ **then**
        Calculate d ; `/* Equation 5.3 */`
        if $d > 0.3$ **then**
            $cal\_value\_prev \leftarrow cal\_value$;
            Calculate cal_value ; `/* Equation 5.5 */`
        end
        Reset $\mathbf{b}^{c}$, $\mathbf{b}^{nc}$ and its flags;
    end
    Push $contact\_state$ to $prediction$;
end

Algorithm 2: Pseudo-code for contact detection using *dynRNN*.

Table C.2: Hyperparameters of convolutional neural network architecture. Every Conv layer is followed by a batch normalization. The outputs of the last two Conv layers within a ResBlock are added and subsequently processed by ReLU. Conv = Convolution, FC = Fully connected, ResBlock = Residual Block

	Layer Type	# Kernel	Kernel Size	Stride	Non-linearity
Conv In	Conv	16	7	1	ReLU
ResBlock 1	Conv	32	7	2	ReLU
	Conv	32	1	2	ReLU after addition
	Conv	32	7	1	ReLU after addition
ResBlock 2	Conv	32	7	2	ReLU
	Conv	32	1	2	ReLU after addition
	Conv	32	7	1	ReLU after addition
ResBlock 3	Conv	64	7	2	ReLU
	Conv	64	1	2	ReLU after addition
	Conv	64	7	1	ReLU after addition
Pooling	Adaptive Avg. Pool	-	-	-	-
Classifier	FC	2	-	-	None

Table C.3: Hyperparameters of convolutional neural network architecture. Every Conv layer is followed by a batch normalization. The outputs of the last two Conv layers within a ResBlock are added and subsequently processed by ReLU. Conv = Convolution, FC = Fully connected, ResBlock = Residual Block

	Layer Type	# Kernel	Kernel Size	Stride	Non-linearity
Conv In	Conv	8	151	1	ReLU
ResBlock 1	Conv	8	21	1	ReLU
	Conv	8	1	1	ReLU after addition
	Conv	8	21	1	ReLU after addition
ResBlock 2	Conv	8	21	1	ReLU
	Conv	8	1	1	ReLU after addition
	Conv	8	21	1	ReLU after addition
ResBlock 3	Conv	16	21	2	ReLU
	Conv	16	1	2	ReLU after addition
	Conv	16	21	1	ReLU after addition
ResBlock 4	Conv	16	21	1	ReLU
	Conv	16	1	1	ReLU after addition
	Conv	16	21	1	ReLU after addition
ResBlock 5	Conv	32	21	2	ReLU
	Conv	32	1	2	ReLU after addition
	Conv	32	21	1	ReLU after addition
ResBlock 6	Conv	32	21	1	ReLU
	Conv	32	1	1	ReLU after addition
	Conv	32	21	1	ReLU after addition
ResBlock 7	Conv	64	21	2	ReLU
	Conv	64	1	2	ReLU after addition
	Conv	64	21	1	ReLU after addition
ResBlock 8	Conv	64	21	1	ReLU
	Conv	64	1	1	ReLU after addition
	Conv	64	21	1	ReLU after addition
Pooling	Adaptive Avg. Pool	-	-	-	-
Classifier 1	FC	8	-	-	ReLu
Classifier 2	FC	2	-	-	None

Table C.4: Model hyperparameters that are optimized during the grid search.

Model	Parameter Name	Values
Extreme Gradient Boosting	Dimensionality reduction	None, 95% PCA
	max_depth	4, 6, 12, 24
	learning_rate	1.0, 0.1
	gamma	0, 0.5, 1
	reg_lambda	0, 10.0, 25.0, 50.0
Support Vector Machine	Dimensionality reduction	None, 95% PCA
	kernel	linear, rbf, poly
	C	0.1, 1, 10
	gamma	scale, auto, 0.1, 1
	degree	2, 3, 4
Logistic Regression	Dimensionality reduction	None, 95% PCA
	penalty	l1, l2, elasticnet
	C	0.1, 1, 10
	solver	saga
	max_iter	10000
	tol	1e-3, 1e-4, 1e-5
k-Nearest Neighbors	Dimensionality reduction	None, 95% PCA
	n_neighbors	3, 5, 7
	weights	uniform, distance
	algorithm	auto, ball_tree, kd_tree, brute
	leaf_size	10, 30, 50
	p	1, 2
Gaussian Naive Bayes	Dimensionality reduction	None, 95% PCA
	var_smoothing	1e-9, 1e-8, 1e-7
Linear Discriminant Analysis Classifier	Dimensionality reduction	None, 95% PCA
Quadratic Discriminant Analysis Classifier	Dimensionality reduction	None, 95% PCA
Perceptron	Dimensionality reduction	None, 95% PCA
	penalty	l1, l2, elasticnet
	alpha	0.0001, 0.001, 0.01
	max_iter	1000, 5000, 10000
	tol	1e-3, 1e-4, 1e-5
	eta0	0.1, 1, 10

Table C.5: Found hyperparameters for the flat and hierarchical classification approaches. N/A indicates that the parameter is not applicable.

Approach	Node	Model	Parameter Name	Value
Flat	root	Support Vector Machine	Dimensionalty reduction	None
			kernel	rbf
			C	10
			gamma	scale
			degree	N/A
Hierarchical	root	Logistic Regression	Dimensionalty reduction	95% PCA
			penalty	l1
			C	0.1
			solver	saga
			max_iter	10000
			tol	0.001
	soft	Support Vector Machine	Dimensionalty reduction	None
			kernel	rbf
			C	10
			gamma	scale
			degree	N/A

Glossary

Bin^1_{hier}	Hierarchical discretization of Young's modulus accounting for manufacturing process variance with level 1
Bin^2_{hier}	Hierarchical discretization of Young's modulus accounting for manufacturing process variance with level 2
Bin_{opt}	Optimal edges for discretization of Young's modulus accounting for manufacturing process variance
Di	Data set $i, \quad i \in \{1, ..., 9\}$, details provided in Table 4.1
Gen_{dev}	Generator that is under development
Gen_{prod}	Generator that is commercially available and in production

Acronyms

ACC	Accuracy
AUROC	Area Under the Receiver Operating Characteristic Curve
BA	Balanced Accuracy
CCA	Counted Contact Areas
CNC	Computer Numerical Control
CNN	Convolutional Neural Network
CNS	Central Nervous System
CPD	Change Point Detection
CWT	Continuous Wavelet Transform
F1	F1-Score
FCN	Fully-Connected Network
GP	Gaussian Process
GRU	Gated Recurrent Unit
hF1	hierarchical F1-Score
hPPV	hierarchical Positive Predictive Value
hTPR	hierarchical True Positive Rate

IV	Information Value
kNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MLNP	Mandatory Leaf Node Prediction
MRI	Magnetic Resonance Imaging
MSS-E	Missed Contact End
MSS-S	Missed Contact Start
NMLNP	Non-Mandatory Leaf Node Prediction
NN	Neural Network
PCA	Principal Component Analysis
PPV	Positive Predictive Value
PVA	Polyvinyl Alcohol
QDA	Quadratic Discriminant Analysis
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SEEG	Stereoelectroencephalography
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
TPR	True Positive Rate
Tr	Training Data
Ts	Test Data
TT-E	Time to Contact End
TT-S	Time to Contact Start
WHO	World Health Organization

Declaration of Generative AI

During the preparation of this thesis the author used ChatGPT and DeepL Write in order to improve readability and clarity of the text. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of this thesis.

Acknowledgments

First and foremost, I want to sincerely thank my supervisor, Prof. Dr. Floris Ernst, for his continuous support of my dissertation and research. His patience, motivation, and immense knowledge have been invaluable. His guidance has been essential throughout my research and writing. I couldn't have asked for a better advisor and mentor.

I also want to thank everyone involved in the KI-SIGS project from the Department of Neurosurgery at the University Hospital of Schleswig-Holstein, Campus Lübeck, and Söring GmbH. Special thanks to Dr. Matteo Bonsanto and Jessica Kren for their clinical expertise and support, and to Dr. Steffen Buschschlüter for his help with the technical aspects and project management. Above all, I am especially grateful to Dr. Daniel Schetelig for his invaluable insights and collaborative spirit, which significantly enriched this work.

I would like to thank my current and former colleagues from the Institute of Robotics and Cognitive Systems for their support, discussions, and coffee breaks. I couldn't have asked for a better team to work with.

I am very thankful to my family and friends for their support and encouragement throughout my studies. My deepest gratitude goes to my wife, Juliane, for her patience, understanding, and love during both the challenging and joyful moments of writing this thesis.

Thank you to everyone who has supported me in any way during this journey. Your support has meant the world to me.