

Dissertation

Deep Learning-based Acoustic Sensing for Medical Applications

Matthias Robert Seibold





Technische Universität München
TUM School of Computation, Information and Technology

Deep Learning-based Acoustic Sensing for Medical Applications

Matthias Robert Seibold

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Julien Gagneur

*Prüfer*innen der
Dissertation:*

1. Prof. Dr. Nassir Navab
2. Assoc. Prof. Dr. Kathleen Denis
3. Prof. Dr. Philipp Fürnstahl

Die Dissertation wurde am 14.11.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 12.04.2023 angenommen.

Matthias Robert Seibold

Deep Learning-based Acoustic Sensing for Medical Applications

Dissertation, Version 1.0

Technische Universität München

TUM School of Computation, Information and Technology

Chair for Computer Aided Medical Procedures & Augmented Reality

Boltzmannstraße 3

85748 Garching bei München

Abstract

Acoustic signals have been utilized by medical professionals for centuries, for example for the diagnostic assessment of human body sounds using a stethoscope, but are rarely used in conventional computer aided diagnostics and surgery. However, acoustic signals have great potential for the development of novel multimodal sensing solutions for medical applications and can provide solutions for problems where conventional systems, such as surgical navigation or medical imaging reach their limits. This dissertation shows that the combination of a highly sensitive sensor technology, advanced signal processing, and powerful cutting-edge analysis methods based on Deep Learning enables the utilization of acoustic signals for the design of easy-to-integrate, non-invasive, radiation-free, and low-cost multimodal sensing systems in computer aided medicine.

The contributions of this work include solutions for unmet clinical problems in diagnostics and surgical interventions which were identified in close collaboration with medical experts. For diagnostics, we present an experimental setup for vibration excitation and deep learning-based vibroacoustic detection of pedicle screw loosening being a post-operative complication after spinal fusion surgery. For interventional use, we propose a system for automated orthopedic drill breakthrough detection based on structure-borne vibroacoustic sensing and a spatio-temporal learning-based framework for the identification of the optimal insertion endpoint for the femoral stem component in Total Hip Arthroplasty using structure-borne hammer blow sounds. All proposed systems were thoroughly evaluated in extensive and realistic human cadaveric experimental setups. To address the problem of limited access to realistic data, which is a common issue in the medical domain due to limited access to the relevant environments as well as strict guidelines and regulations, a novel data augmentation method based on a conditional generative adversarial network is proposed which is able to generate realistic synthetic samples from a learned dataset distribution to improve the performance of medical acoustic sensing systems.

Our results proof that automated decision and support systems based on acoustic sensing have great potential for the development of new multimodal sensing paradigms for a wide range of applications in medical diagnostics, interventions, and the analysis of surgical workflows. Acoustic sensing systems can be utilized to complement existing computer aided surgery, surgical guidance, and decision support systems and provide information beyond the limits of established methods such as surgical navigation systems or medical imaging.

Zusammenfassung

Akustische Signale werden von Mediziner*innen seit Jahrhunderten verwendet, z. B. zur diagnostischen Beurteilung menschlicher Körpergeräusche mit Hilfe eines Stethoskops, finden aber nur selten in der herkömmlichen computergestützten Diagnostik und Chirurgie Anwendung. Die Nutzung von akustischen Signalen hat jedoch großes Potenzial für die Entwicklung neuartiger multimodaler Sensorlösungen in der Medizin und kann Lösungen für Probleme bieten, bei denen herkömmliche Systeme wie chirurgische Navigation oder medizinische Bildgebung an ihre Grenzen stoßen. Diese Dissertation zeigt, dass die Kombination von hochsensibler Sensortechnologie, fortschrittlicher Signalverarbeitung und leistungsstarker, auf Deep Learning basierender Analysemethoden die Nutzung akustischer Signale für die Entwicklung einfach zu integrierender, nicht-invasiver, strahlungsfreier und kostengünstiger multimodaler Sensorsysteme in der computergestützten Medizin ermöglicht.

Die Beiträge dieser Arbeit umfassen Lösungen für ungelöste klinische Probleme in der Diagnostik und Chirurgie, die in enger Zusammenarbeit mit medizinischen Expert*innen identifiziert wurden. Als diagnostische Anwendung stellen wir einen experimentellen Aufbau für die Schwingungsanregung und die vibroakustische Erkennung von Pedikelschraubenlockerungen, eine postoperative Komplikation nach Wirbelsäulenfusionsoperationen, vor. Für die interventionelle Anwendung schlagen wir ein System zur automatischen Erkennung von Bohrerdurchbrüchen in der Orthopädie, sowie ein auf spatiotemporal Learning basierendes System zur Identifizierung des optimalen Einführungsendpunkts der Femurschaftkomponente in der Hüfttotalendoprothetik vor. Alle, im Rahmen dieser Dissertation entwickelten Systeme wurden in umfangreichen und realistischen Versuchen mit menschlichen Kadavern evaluiert. Da im medizinischen Bereich aufgrund des begrenzten Zugangs zur relevanten Umgebung sowie strenger Richtlinien und Vorschriften der Zugang zu realistischen Daten oft begrenzt ist, wird eine neuartige Methode zur Vergrößerung eines Datensatzes auf der Grundlage eines Generative Adversarial Networks vorgestellt, die die Genauigkeit medizinischer Acoustic Sensing Systeme zu verbessern kann.

Unsere Ergebnisse belegen, dass automatisierte Unterstützungssysteme auf der Basis akustischer Sensorik großes Potenzial für die Entwicklung neuer multimodaler Sensorik-Paradigmen für eine Vielzahl von Anwendungen in der medizinischen Diagnostik, für Interventionen und bei der Analyse chirurgischer Arbeitsabläufe haben. Akustische Sensorsysteme können zur Ergänzung bestehender computergestützter Chirurgie- und medizinischer Entscheidungsunterstützungssysteme eingesetzt werden und Informationen liefern, die über die Grenzen etablierter Methoden wie chirurgischer Navigationssysteme oder medizinischer Bildgebung hinausgehen.

Acknowledgments

First of all, I want to thank my doctoral supervisor Prof. Dr. Nassir Navab for giving me the opportunity to pursue my PhD in the first place and work in the environment of the Research in Orthopedic Computer Science group of Prof. Dr. Philipp Frnstahl at Balgrist University Hospital in close collaboration with the Computer Aided Medical Procedures (CAMP) chair at TU Munich. I'm thankful that he always believed in me, gave me the freedom to pursue research in the field of acoustics that I was most interested in, and gave me many times the right advice and valuable feedback in scientific questions. He furthermore initiated the Medical Augmented Reality Summer School (MARSS) and gave me the chance to organize the event two times which was an exciting experience and connected me with a lot of inspiring people from all over the world.

I thank Prof. Dr. Philipp Frnstahl who welcomed me in his group in Zurich, who greatly supported me, was always helpful and gave feedback when I needed it, and was responsive for all questions related to scientific and organizational matters arising during my daily work. I'm very happy that I was able to do my PhD in the extraordinary research environment of the Balgrist Campus and that he gave me the freedom to follow my interest and granted me responsibility and trust which were the main factors for my professional and personal development during the time of my PhD.

I furthermore want to thank Prof. Dr. med. Mazda Farshad who enabled me to pursue my doctoral studies at Balgrist University Hospital in the context of the HMZ flagship project "SURGENT". Prof. Farshad's excitement and interest for science creates a thriving research environment and facilitates the exchange between engineers and clinicians which is very special. Here, I want to thank in particular my clinical collaborators, namely Dr. med. Armando Hoch and Dr. med. Christoph J. Laux with whom I spent many days performing experiments and discussing exciting ideas on the side of their tightly packed clinical schedule.

Big thanks to all the researchers, engineers, clinicians, and administrative staff from the ROCS group for supporting my research through guidance, help, and discussions and the colleagues at the Balgrist Campus for the fruitful collaborations and for providing a pleasant working environment. Thanks to all people at the CAMP chair at TU Munich for widening my scientific horizon, good times at CAMPing, and especially to the NARVIS team around Dr. Ulrich Eck for weekly discussions, feedback, and support for organizing MARSS. I also would like to thank Daniel Ostler and the people at MITI for introducing me to the whole research environment at TUM and sparking my interest in the field of Acoustic Sensing. Furthermore, I want to thank

Navid Navab from Topological Media Lab at Concordia University, Montreal, for exciting and inspiring discussions around the topic of audio, acoustic sensing, and sonification.

Last but not least, I want to thank my family, especially my parents Silvia and Gerhard, but also their partners, my brother, my grandparents and relatives, for their love and care, for supporting me throughout my whole life and enabling me to follow my dreams. Also my friends in Zurich, Munich, Günzburg, and many other places on this planet have a great part in who and where I am today and I am deeply grateful for their love and friendship.

Contents

List of Authored and Co-authored Publications	1
I Introduction	3
1 Introduction	5
1.1 Thesis Outline	5
1.2 Motivation	5
1.2.1 Acoustic Signals in Medicine	5
1.2.2 Clinical Motivation	7
1.3 Technical Background and Related Work	8
1.3.1 Acoustic Sensing for Medical Diagnostics and Monitoring	8
1.3.2 Interventional Acoustic Sensing	9
1.3.3 The Rise of Deep Learning for Audio Signal Processing	11
1.4 Thesis Objective and Contributions	13
II Methodology and Contributions	15
2 Audio Signal Processing	17
2.1 Acoustic Signals	17
2.2 Processing Acoustic Signals	19
2.2.1 Audio Capturing and Preprocessing	19
2.2.2 Audio Features	20
2.2.3 Spectrograms	22
3 Machine Learning-based Acoustic Sensing	25
3.1 Classical Approaches	25
3.2 Spectrogram-based Audio Processing	25
3.2.1 Contribution (Interventional): Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery (Scientific Reports 2021)	26
3.2.2 Contribution (Diagnostic): A new sensing paradigm for the vibroacoustic detection of pedicle screw loosening (Not peer-reviewed)	39
3.3 Spectrogram-based Spatio-temporal Learning	51
3.3.1 Contribution (Interventional): Acoustic-Based Spatio-Temporal Learning for Press-Fit Evaluation of Femoral Stem Implants (MICCAI 2021)	51
3.4 Approaches based on Raw Waveforms	63
3.5 Data Augmentation	63

3.5.1	Contribution: Conditional Generative Data Augmentation for Clinical Audio Datasets (MICCAI 2022)	63
3.5.2	Contribution: Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data (MICAD 2022)	75
III	Conclusions and Outlook	87
4	Summary of Findings	89
5	Future Directions	93
IV	Appendix	97
A	Abstracts of Publications not Discussed in this Thesis	99
	Bibliography	105
	List of Figures	115
	List of Tables	117

List of Authored and Co-authored Publications

2022

- [103] Marine Y. Shao, Tamara Vagg, **Matthias Seibold**, Mitchell Doughty. "Towards a Low-Cost Monitor-Based Augmented Reality Training Platform for At-Home Ultrasound Skill Development". *Journal of Imaging*, 8 (11), 305, 2022.
- [62] Mane Margaryan*, **Matthias Seibold***, Indu Joshi, Mazda Farshad, Philipp Fürnstahl, Nassir Navab. "Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data". *arXiv preprint, arXiv:2211.02874, accepted for Medical Imaging and Computer-Aided Diagnosis 2022, Leicester, UK, 2022. (*equal contribution)*
- [101] **Matthias Seibold**, Bastian Sigrist, Tobias Götschi, Jonas Widmer, Sandro Hodel, Mazda Farshad, Nassir Navab, Philipp Fürnstahl, Christoph J. Laux. "A new sensing paradigm for the vibroacoustic detection of pedicle screw loosening". *arXiv preprint, arXiv:2210.16170, 2022.*
- [106] Tianyu Song, Michael Sommersperger, The Anh Baran, **Matthias Seibold**, and Nassir Navab. "HAPPY: Hip Arthroscopy Portal Placement Using Augmented Reality". *Journal of Imaging*, 8(11), 302, 2022.
- [67] Sasan Matinfar, Mehrdad Salehi, Daniel Suter, **Matthias Seibold**, Navid Navab, Shervin Dehghani, Florian Wanivenhaus, Philipp Fürnstahl, Mazda Farshad, Nassir Navab. "Sonification as a Reliable Alternative to Conventional Visual Surgical Navigation". *arXiv preprint, arXiv:2206.15291, 2022.*
- [98] **Matthias Seibold**, Armando Hoch, Mazda Farshad, Nassir Navab, and Philipp Fürnstahl. "Conditional Generative Data Augmentation for Clinical Audio Datasets". *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI) 2022, Singapore, Singapore, 2022.*

2021

- [99] **Matthias Seibold**, Armando Hoch, Daniel Suter, Mazda Farshad, Patrick O. Zingg, Nassir Navab, and Philipp Fürnstahl. "Acoustic-Based Spatio-Temporal Learning for Press-Fit Evaluation of Femoral Stem Implants". *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI) 2021, Strasbourg, France, 2021*.
- [36] Jonas Hein*, **Matthias Seibold***, Federica Bogo, Marc Pollefeys, Mazda Farshad, Philipp Fürnstahl, Nassir Navab. "Towards markerless surgical tool and hand pose estimation". *International Journal of Computer Assisted Radiology and Surgery*, 16, pp. 799–808, 2021. (*equal contribution)
- [100] **Matthias Seibold**, Steven Maurer, Armando Hoch, Patrick O. Zingg, Mazda Farshad, Nassir Navab, and Philipp Fürnstahl. "Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery". *Scientific Reports*, 11 (1), 2021.

2020

- [78] Daniel Ostler*, **Matthias Seibold***, Jonas Fuchtmann, Nicole Samm, Hubertus Feussner, Dirk Wilhelm, Nassir Navab. "Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions". *International Journal of Computer Assisted Radiology and Surgery*, 15, pp. 771–779, 2020. (*equal contribution)¹
- [84] Mustafa Haiderbhai, Sergio Ledesma, Sing Chun Lee, **Matthias Seibold**, Philipp Fürnstahl, Nassir Navab, Pascal Fallavollita. "pix2xray: converting RGB images into X-rays using generative adversarial networks". *International Journal of Computer Assisted Radiology and Surgery*, 15, pp. 973-980, 2020.
- [54] Sing Chun Lee*, **Matthias Seibold***, Philipp Fürnstahl, Mazda Farshad, Nassir Navab. "Pivot calibration concept for sensor attached mobile c-arms", *In Proceedings of the Image-Guided Procedures, Robotic Interventions, and Modeling 2020, Houston, TX, USA, 2020*. (*equal contribution)

2019

- [66] Sasan Matinfar, Thomas Hermann, **Matthias Seibold**, Philipp Fürnstahl, Mazda Farshad, Nassir Navab, "Sonification for Process Monitoring in Highly Sensitive Surgical Tasks". *In Proceedings of the Nordic Sound and Music Computing Conference 2019 (Nordic SMC 2019), Stockholm, Sweden, 2019*.

¹In this work, the results of the author's Master Thesis have been published. The findings are presented in the context of this dissertation as a first preliminary work in the field of Deep Learning-based Acoustic Sensing for Medical Applications.

Part I

Introduction

Introduction

1.1 Thesis Outline

In the following paragraphs, a brief outline of this dissertation is presented.

Chapter 1 introduces the concept of using acoustic signals in the context of medical history and in current clinical practice. In this scope, the clinical motivation of this work is illustrated. An overview about the technical background and related work is given and the objective of the present thesis based on the clinical motivation and technical state-of-the-art is presented.

Chapter 2 explains the technical background of acoustic signals, from signal capturing and digitization to signal processing and feature extraction methods for audio signals.

Chapter 3 presents an overview of learning-based methods for medical acoustic sensing, starting from classical approaches using handcrafted features to the author's contributions in state-of-the-art frame-based and spatio-temporal learning-based acoustic sensing solutions in the context of computer assisted orthopedic surgery and diagnosis. Furthermore, raw waveform-based approaches are briefly discussed and a contribution in the field of data augmentation for clinical audio data is presented.

Chapter 4 summarizes the findings of this dissertation. The thesis is concluded with an outlook for future research directions in **Chapter 5**.

1.2 Motivation

1.2.1 Acoustic Signals in Medicine

The usage of acoustic signals has a long history in medicine. *Immediate auscultation*, a technique which is characterized by placing the ear directly onto the patients's chest to detect abnormal chest sounds was first described in ancient times by Hippocrates (approx. 460 - 370 BC) [56]. In 1816, a french physician called Rene Theophile Hyacinthe Laennec invented the *stethoscope* by rolling 24 papers in the shape of a cone and placing the instrument on the patient's chest to transmit structure-borne chest sounds to the examining physician's ear. He furthermore published an article in 1822, describing techniques for the auscultation of lung and heart sounds, e.g. for diseases such as tuberculosis [50] and improved the design of the stethoscope as illustrated in figure 1.1. Since these early usages of acoustic signals in medical diagnosis, the assessment, description, and analysis of human body sounds has become an



Fig. 1.1. The Laennec stethoscope. The examining physician would place the wooden stethoscope on the patient's chest and their ear on the other end of the tubular structure to listen to heart and lung sounds for abnormalities.¹

integral part for the training and education of medical professionals. The modern version of the stethoscope is an important tool for doctors all around the world and closely associated with the medical profession. Especially for cardiac and pulmonary applications, auscultation is a widely adopted method for the examination of patients. Also in other use cases, such as the examination of bowel sounds or the assessment of cartilage degeneration in bone joints, auscultation has proven to be a reliable medical tool.

Also in the operating theatre, acoustic signals are always present which includes sounds generated by the surgeon's interaction with the patient, such as diathermy, drilling, or hammering, continuous signals from surgical devices such as heart monitors, alarms and notification signals, as well as the communication of the surgical staff during interventions. These acoustic signals contain highly dense information about the current state of the procedure and characteristic surgical events. Especially in orthopedics, where drilling, hammering, chiseling, sawing, and other mechanical interactions with the anatomy generate characteristic noise, experienced surgeons report that together with visual and haptic cues, they are able to infer additional information about the surgical action from acoustic signals. Examples include the assessment of the seating of orthopedic implants during insertion, the differentiation between different tissue types, such as cancellous bone, cortical bone, and cartilage during chiseling, sawing, or drilling, and the assessment of biomechanical properties of the bone and degenerative diseases, such as osteoporosis, e.g. during bone screw insertion. In addition, also interaction with soft tissues such as coagulation, needle insertion or cutting with sharp tools creates distinct audible noise or structure-borne vibrations.

Through the advances in sensor hardware, medical acoustic signals can nowadays be captured using highly sensitive and easy-to-integrate sensors in an air-borne and/or structure-borne manner. By using high-quality airborne microphones, it is possible to capture room sounds (using microphones with sphere or cardioid polar patterns) or sound sources (using directed microphones). Contact microphones allow the acquisition of high quality structure-borne acoustic signals from tools or directly attached to the patient skin or anatomy. The following

¹retrieved from [https://de.wikipedia.org/wiki/Ren%C3%A9_Laennec#/media/File:Laennecs_stethoscope,_c_1820._\(9660576833\).jpg](https://de.wikipedia.org/wiki/Ren%C3%A9_Laennec#/media/File:Laennecs_stethoscope,_c_1820._(9660576833).jpg) on August 11th, 2022, released under Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0) [background removed]

section gives an overview about challenges for clinical decision making in clinical diagnosis and interventions and illustrates how acoustic sensing can be used to address these challenges.

1.2.2 Clinical Motivation

Clinical decision making is a challenging task, in both the diagnostic and intraoperative context. In diagnostics, decision making is often based on the examination of medical images such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) or planar radiographs (X-Rays) which involves the assessment and interpretation of the imaging data by medical professionals. This process is however not flawless and errors in decision making can not be fully excluded [49]. While in current clinical practice, examination and diagnosis is moving towards objective decision-making procedures involving medical tests and medical imaging, medical professionals utilize subjective senses, such as vision, hearing (as described above) or smell, which still plays a crucial role for the interpretation of images, test results, the monitoring of a patient during hospitalization, and the final decision making in patient diagnosis [64]. Therefore, multimodal machine sensing can play an important role in supporting medical diagnostics and has great potential to improve the diagnostic outcome by analyzing additional sensor data for an informed decision.

Intraoperative decision making is a complex process and involves intuitive and analytic aspects, the assessment of the patient risk, the individual anatomical patient characteristics, and many more factors [82]. One example for challenges in intraoperative decision making is tissue differentiation which is especially demanding in minimally invasive procedures due to lack of haptic feedback, visual occlusion, limited access, and the similar visual appearance of different types of tissue [65]. Also in open surgery, tissue differentiation, e.g. for the differentiation between tumor and healthy tissue or the navigation of power tools in occluded and hard-to-reach areas of the anatomy, is an important part of an intervention. Computer assisted surgical navigation has been introduced to provide surgical guidance by determining the position of the patient anatomy and the surgical tools in 3D space, usually through optical and electromagnetic tracking systems, or robotic devices [42]. However, there are several limitations of surgical navigation systems, for example line-of-sight issues, the requirement of offline or online registration procedures to align the patient anatomy with the preoperative plan, the lack of capturing semantic information from the area of operation, or the estimation of tissue deformation in relation to the preoperative plan. Furthermore, it is impossible to assess certain surgical measures, such as the seating and press-fit of an implant in the bone, only with optical systems. For instance, the optimal insertion endpoint, e.g. for a femoral hip implant, is still dependent on the expertise of the operating surgeon [4]. Surgeons rely on multiple senses to interpret and interact with the world which gives them the ability to combine information from different sources to make a decision. Inspired by the multimodal sensing of humans, there is an increasing body of research in the direction of multimodal sensing for improved surgical guidance and autonomous surgical robotics. In this context, advanced methods, for example hyperspectral imaging [26] or vibroacoustic sensing [19] have great potential to capture valuable additional information about the intervention.

Acoustic signals, which have been used by physicians and medical professionals since ancient times, provide an easy-to-integrate, non-invasive, radiation-free, and low cost sensing modality.

They contain dense and high quality information which can be used for multimodal sensing approaches in medical diagnostics and interventions. The advances in signal processing and pattern recognition enable the design of automated decision making systems which have the potential to support and inform physicians while performing surgical actions, prevent surgical and diagnostic errors, detect adverse events, provide additional information, e.g. for tissue differentiation in computer aided surgery, or analyze surgical workflow. Furthermore, medical acoustic sensing systems can be a valuable tool to support medical diagnostics in applications such as chest diseases, cartilage degeneration, or bone joint implant monitoring. The following sections give an overview about the state-of-the-art of applications for acoustic signals in diagnostic and monitoring applications, as well as surgical interventions.

1.3 Technical Background and Related Work

1.3.1 Acoustic Sensing for Medical Diagnostics and Monitoring

There is a wide variety of applications for acoustic sensing in medical diagnostics. The following paragraphs give an overview of the state-of-the-art of acoustics in medical diagnosis. As described in section 1.2.1, the first applications of acoustics for the diagnosis of chest diseases originate from ancient times, where the examining physician placed the ear directly on the patient's chest to listen to abnormalities in lung and heart noise. In the modern times, the field of automated chest sound analysis is widely covered. Example applications are a phonocardiogram automatic classification model based on a CNN classifier which is able to assist physicians in the diagnosis of heart sounds [119]. A RNN-based method for the analysis of heart sounds was proposed by Yang et al. [120]. Marshall et al. published an algorithm which allows non-specialists to screen for pulmonary fibrosis [63]. A smart stethoscope which analyzes respiratory sounds using a deep learning-based classifier was proposed by Ma et al. [60] in 2019. Further related applications include voice pathology classification using speech recordings of patients, e.g. a system proposed by Miliaresi et al. [69]. In the COVID-19 pandemic, systems for the automated analysis of cough sounds for characteristics of the COVID-19 virus have been proposed [51].

The Assessment of orthopedic implants based on acoustic sensing has the advantages of being non-destructive and radiation-free and can be used for early diagnostics. The main applications in the literature are the analysis of hip and knee replacement implants [44]. Schwarzkopf et al. analyzed data of different types of knee implants using a handheld measurement system which revealed correlations of the acoustic signature to the implant status and time from implantation [97]. Rodgers et al. characterized the squeaking of hard-on-hard bearing surface combinations by acoustic emission analysis of Total Hip Arthroplasty (THA) implants [90]. A monitoring system to assess the wear of THA implants was developed by Fitzpatrick et al., who compared the frequency characteristics of in-vivo and in-vitro signals [27]. Ewald et al. developed a measurement prototype and simulator-based experimental setup to detect loosening of THA implants using an acoustic sensor [25]. Arami et al. applied harmonic

vibration to the tibia and captured the resulting vibrations at the surface of a total knee replacement implant using an accelerometer-based sensor [6].

Another research area that is covered in the literature is the field of *Vibroarthrography* (VAG), a technique that uses vibroacoustic signals captured with accelerometers or contact microphones on the patient's skin to record movement noise of the joint and determine various stages of cartilage degeneration, usually in the knee joint. Frank et al. published a paper in 1990 in which they analyzed knee joint acoustic signals and described their nature and diagnostic potentials [28]. Krishnan et al. conducted a clinical study which showed that VAG has potentials as a diagnostic tool for the screening of chondromalacia patella [47]. An automated algorithm based on handcrafted features and a FCNN to classify knee joint acoustic signals was proposed by Kim et al. [45] which showed promising performance for the diagnosis of diseases such as osteoarthritis. Further research projects involved different handcrafted features, employing automated feature selection techniques [73], wavelet filter bank features [104], or frequency features from multiple sensors [7], using classical supervised learning techniques such as SVMs or RFs.

The monitoring of the elderly in their home, e.g. for fall and other acoustic event detection, has great potential to minimize hospitalization costs, reduce the need for nursing staff, and enable the patients to live in their own home, therefore improving the quality of life in old age. For this type of monitoring application, ambient assisted living systems have been introduced which mainly rely on the analysis of acoustic signals recorded from environmental microphones. Example systems include a concept for the monitoring of an entire flat which uses a two-stage acoustic event classification approach that is able to detect events related to dangerous health conditions and was proposed by Navarro et al. [75]. Another example was proposed by Ghayvat et al., who developed a system involving a microphone prototype and a detection method based on a CNN for at-home monitoring of elderly [29].

Acoustic signal analysis is also an integral part of wearables for health monitoring. Applications include the precise monitoring of the daily food intake of a person [8], the mechano-acoustic monitoring of cardiopulmonary signals [35], or the discrimination of fetal movement during pregnancy [52].

1.3.2 Interventional Acoustic Sensing

In the context of clinical interventions, acoustic signals provide meaningful and highly dense information about surgical actions, the process, and state. The following paragraph gives an overview of how acoustic signals are used in interventional applications in the current state-of-the-art and research literature.

Especially in orthopedic interventions, where surgical interactions with hard tissues (bone and cartilage) such as hammering, drilling, sawing, reaming, chiseling, etc. cause characteristic sounds, medical acoustic sensing systems have been researched for decades. For example in surgical drilling, which is one of the most frequent tasks in orthopedic surgery and is part of many orthopedic interventions, a study by Praamsma et al. showed, that surgeons implicitly utilize the sounds generated by surgical drilling to guide the drilling motion [86]. Therefore,

the automated analysis of drilling sounds is a promising research direction which has been extensively covered in the literature. Boesnach et al. developed a method to analyze drill sounds in spine surgery by applying FCNNs, Support Vector Machines (SVM) and Hidden Markov Models (HMM) to spectral density estimates [9]. Shine et al. recorded acoustic signals in a cadaveric experiment during otologic drill and analyzed the spectral energy of different frequency bands. They could show that the acoustic signature of drill burr-bone interface differs between thick calvarial bone and thin tegmen bone [105]. Furthermore, systems have been developed that leverage drill sounds to perform automated drill state detection, i.e. performing tissue differentiation based on recorded drilling noise. Sun et al. [109] proposed a system based on power spectral density estimates to perform drill state detection in an experiment with porcine specimens. A method for the differentiation between cortical and cancellous bone drilling samples in an experiment with bovine specimens which is based on Short-term Fourier Transform (STFT) features combined with SVM, HMM and Random Forest (RF) classifiers was proposed by Zakeri et al. [122, 123]. Guan et al. proposed a system to control the drilling of pedicle screw based on handcrafted features computed from acoustic signals and a FCNN [34]. A system for the detection of drill breakthrough in a rat skull was developed by Pohl et al. [85] who utilized spectral density features and a SVM classifier. Torun et al. developed a method based on four frequency-based handcrafted features and a FCNN for the task of drill breakthrough detection in an experimental setup [112].

Also for surgical milling, in which the bone is cut with a rotary tool which is, in contrast to drilling, moved orthogonal to the rotation axis, process monitoring using acoustic signals has been covered in the literature. Dai et al. developed a method based on lifting wavelet packet transform to compute wavelet energy which serves as input for a SVM classifier to differentiate between cancellous and cortical bone [18], as well as between vertebra, spinal cord and muscle tissue [20]. A similar method was developed by Ying et al., who developed a system using acoustic signals to distinguish between cortical bone, cancellous bone and idle state in lumbar laminectomy [121].

Another application for acoustic sensing technology is the monitoring of implant insertion in orthopedic surgery. Morohashi et al. identified correlations of hammer blow sounds to complications during the insertion of femoral stem implants in THA [71]. The research direction was continued by Oberst et al., who performed a nonlinear timeseries analysis with the impulse response of air-borne hammer blow sounds [76] and Tijou et al. [111] and Dubory et al. [23], who developed a sensorized hammer and found correlations to implant displacement in time-domain and peak-based features computed from the captured structure-borne vibrations. Goossens et al. computed a set of handcrafted features based on the energy distribution in the frequency spectrum of hammer blow sounds captured with an air-borne microphone and validated the approach in a real surgery setting [32]. They furthermore observed a resonance frequency shift during the insertion process of the acetabular hip implant component in an in-vitro experiment [31]. Wei et al. found certain peak frequencies to statistically increase with the insertion depth of the femoral component of THA implants in a lab experiment using a custom bone phantom [117].

For the application of interventions involving needles, promising results have been obtained using acoustic signal analysis. A prototype for the identification of tissue penetration based on time-varying auto-regressive (TV-AR) was proposed by Illanes et al. [38]. They showed

that the structure-borne audio captured by a sensor placed on the distal end of the needle could provide complementary information for intraoperative guidance with different setups and interventional devices [39]. Their research group furthermore proposed the usage of a similar system for Veress needle insertion in minimally invasive surgery (MIS) [96] and for the differentiation of varying degrees of cartilage degeneration in arthroscopy by palpating the cartilage surface with a Veress needle [108].

While in conventional surgery physicians rely on visual, auditory, and haptic cues for information retrieval, the haptic sensation is strongly reduced in robotic surgery which, on the other hand, offers many advantages such as tremor compensation, micro-precision through gear reduction, or teleoperation [79]. Chen et al. proposed a prototype to analyze structure-borne vibrations caused by the interaction of a robotic grasper with different biological texture samples. In their experiments, they could show a correlation between spectral information and grasped textures. [14]

Soft tissue differentiation is challenging because structures appear visually similar and critical structures are not clearly distinguishable. For the application in burn surgery for necrotic tissue removal, a system for tissue differentiation was proposed by Nahen et al., who used acoustic signals generated by the application of an Er:YAG laser to vital and necrotic tissues [72]. Alperovich et al. developed a system for the differentiation of acoustic signals generated by ultraviolet laser ablation in vascular procedures using Mel-frequency cepstral coefficients and SVM and FCNN classifiers [5]. The limited visual access of a laparoscope makes the tissue differentiation in minimally invasive surgery even more difficult. To tackle this problem, the author developed a system in his Master Thesis to differentiate between muscle, fat, fascia and liver tissue by analyzing cauterization sounds from within the surgical operation area, a work which was published as a journal paper during the course of the PhD studies [78]. This first proof-of-concept study serves as the basis for the works discussed in the course of this thesis.

In addition to the above mentioned interventional use-cases and applications of acoustic signals for medical application, acoustic signals have also been utilized as a complementary source of information for the automated analysis of surgical workflow [116].

1.3.3 The Rise of Deep Learning for Audio Signal Processing

The year 2012, when a deep learning-based method by Krizhevsky et al. called AlexNet [48] won the prestigious ImageNet competition [92] and outperformed the competition by a top-5 error margin of 10.8%, is commonly considered as the breakthrough of deep learning. This success was made possible by novel neural network architectures with large numbers of parameters trained on huge data sets utilizing powerful dedicated parallel computing hardware. Since then, deep learning methods have been applied to a plethora of problems in various research fields. Popular architectural designs for deep learning models are feedforward neural networks [70], Convolutional Neural Networks (CNNs) [53], recurrent neural networks (RNNs) such as long short-term memory (LSTM) [37] or gated recurrent units (GRU) [16], and more recently attention- [114] and graph-based [10] architectures. The training of deep neural networks is based on gradient computation using the backpropagation algorithm [91]

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The model learns to map inputs to outputs using input-output pairs of data.	The models learns patterns from unmapped data.	An agent learns to maximize a reward by taking actions in an environment.
Problems	Classification, Regression	Association, Clustering, Synthetic Data Generation	Reward-based problems
Examples	Image Classification, Object Detection, Pose Estimation	Anomaly Detection, Data generation (Autoencoders, GANs), Recommendation Systems	Autonomous Driving, Recommendation Systems, Games

Tab. 1.1. The three types of machine learning including definitions, typical problems and example applications.

and parameter optimization such as the infamous *Gradient Descent* algorithm, which was first proposed by the french mathematician Augustin-Louis Cauchy in 1847. Many different variants of deep neural networks have been proposed in the literature which can be categorized into three types of machine learning, *Supervised Learning*, *Unsupervised Learning*, and *Reinforcement Learning*, which are illustrated in Table 1.1. This thesis covers supervised learning methods for the design of automated medical acoustic decision support systems, as well as unsupervised learning aspects of synthetic data generation for data augmentation. Reinforcement learning, which can be utilized in the medical domain for example for surgery planning [3], is included in table 1.1 for completeness, but not discussed in the scope of the present work. Until now, deep learning systems have been continuously improved and specialized architectures have been developed to tackle complex problems. State-of-the-art models achieve astonishing results in image classification with *Vision Transformers* [46], novel view synthesis for complex scenes using *Neural Radiance Fields (NeRF)* [68], synthetic image generation with *DALL-E-2* [89] or *Imagen* [93], text generation with *GPT-3* [11], or protein folding with *AlphaFold* [43].

Besides advances in the fields such as Computer Vision, Natural Language Processing, Autonomous Driving, or Recommender Systems, deep learning has also revolutionized the field of acoustics and audio signal processing [87]. For example in the field of human speech recognition, the established methods, which were usually based on classical Machine Learning techniques such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), or simple fully connected neural networks (FCNNs) were outperformed by large-scale deep learning models [74]. A similar development can be observed in the fields of Acoustic Scene Classification [1] and Environmental Sound Event Recognition [12]. In comparison to classical approaches such as GMMs, HMMs, or FCNNs, deep learning-based methods do not rely on the computation of selected handcrafted features, but are able to learn the feature extraction implicitly by using raw waveforms or audio spectrograms as input [87]. In addition, it is not only possible to use recorded acoustic signals as input for learning-based systems, but also use the power of deep learning to synthesize computer-generated audio for realistic text-to-speech

(TTS) conversion like *WaveNet* [77] or for music synthesis [40]. Acoustic emission monitoring based on neural networks is also employed for failure prediction of industrial equipment by analyzing the mechanical noise and vibrations of engines, bearings, valves [88] or drill bits [115]. Furthermore, deep-learning based acoustic sensing has been employed for first medical applications, such as COVID-19 through cough sound detection [51], anomaly detection in heart [120] and lung [60] sounds, or voice pathology classification [69].

1.4 Thesis Objective and Contributions

As discussed in section 1.2.2, there is an unmet clinical need to complement and replace established processes and systems for clinical decision making with multimodal sensing data. To this end, as described in sections 1.3.1 and 1.3.2, acoustic signals provide an easy-to-integrate, non-invasive, radiation-free, and low-cost sensing modality and have already been successfully employed for a variety of diagnostic medical applications, e.g. in chest medicine, orthopedic implant assessment, and the monitoring of cartilage degeneration, as well as for patient monitoring and surveillance. The combination of high-quality signal acquisition, advanced signal processing, and state-of-the-art deep learning-based analysis methods, enables the utilization of air- and structure-borne acoustic signals in the medical context to design automated decision making systems and develop surgical and diagnostic guidance and decision support solutions. Therefore the objective of this dissertation is to find novel ways for using acoustic signals to improve clinical decision making, develop clinically feasible hardware setups that allow the translation to clinical applications, and design novel state-of-the-art deep learning-based methods and reliable automated acoustic sensing solutions for orthopedic applications. The present thesis explores the field of deep learning-based acoustic sensing for medical applications and presents novel and specialized methods for the automated analysis of acoustic signals for multiple medical applications in interventional and diagnostic scenarios which have been developed in the course of the author's PhD studies.

The conceptualization of this project started with the Master Thesis of the author which was conducted at the research group Minimally Invasive Interdisciplinary Therapeutical Intervention (MITI) at the university hospital "Rechts der Isar" in Munich based on discussions with Navid Navab from the Topological Media Lab at Concordia University Montreal, Canada. The work was published and received the best paper award at the IPCAI conference. The abstract of the published paper is included in the appendix of this work in section A. Within this work, a system for the differentiation of different types of soft tissue by analyzing coagulation sounds in minimally invasive surgery was developed [78]. In the context of the present thesis, which was conducted in cooperation with Balgrist University Hospital, University of Zurich, the concept of acoustic sensing in medical applications was advanced, extended, and improved towards unmet clinical needs in orthopedic surgery and diagnostics. Hereby, a goal was to develop signal processing pipelines and state-of-the-art deep learning methods customized on specialized use cases in clinical orthopedics and incorporate the individual challenges in the medical domain such as task-specific algorithm design and the handling of small data set sizes and imbalanced data sets. The system development and data collection was performed in realistic environments, including extensive human cadaver experiments, biomechanical testing setups, and data acquisition in the real-world operating room. In addition, practical concepts for capturing high-quality audio signals in diagnostic and surgical environments

have been developed, including custom, modular structure-borne microphones to capture vibroacoustic signals directly from the patient anatomy or surgical tools, as well as air-borne recording setups to capture room audio or acoustic signals from the area of operation during surgery.

Within the course of the present thesis, contributions for multiple acoustic sensing applications in the orthopedic domain have been achieved and published in high-quality scientific journals and conferences. The thesis presents multiple solutions for clinical applications where established computer aided surgery and diagnosis systems such as surgical navigation systems or medical imaging reach their limits. In the first presented application (section 3.2.1), the purpose of the system is to provide surgical error prevention for hand-held drilling which has to be realized with as minimal latency as possible to minimize the risk of drill breakthrough and therefore potential harm to soft tissue. The second presented paper (section 3.2.2) proposes a novel sensing paradigm to assess the hold of pedicle screws², a clinical challenge where established diagnosis methods such as medical imaging have been shown to fail in a substantial amount of cases. The third work, described in section 3.3.1 introduces a novel system based on spatio-temporal learning which analyzes structure-borne hammer blow sounds captured with a contact microphone from the inserter tool during THA surgery to assess the press-fit during femoral stem insertion. The system enables the identification of the optimal insertion endpoint of the femoral stem component, a measure which is not obtainable neither with medical imaging, nor with conventional surgical navigation systems. Furthermore, a novel data augmentation strategy based on a generative model for clinical audio data sets is presented in section 3.5.1 which is able to improve the performance of medical acoustic sensing systems and tackle the problem of limited data set sizes in the medical domain. The method was subsequently improved and extended which is described in section 3.5.2.

²This work did not undergo peer-review and is not relevant for the grading of this dissertation.

Part II

Methodology and Contributions

Audio Signal Processing

2.1 Acoustic Signals

” *Acoustics is defined as the science that deals with the production, control, transmission, reception, and effects of sound.*

— Marriam-Webster Dictionary

Acoustics¹ is a vast research field and covers many aspects related to acoustic signals, from music theory to architectural acoustics, from infrasound which is used in sonar systems to ultrasound used for medical imaging. The wheel of acoustics, as illustrated in Figure 2.1 gives an overview of the scope of acoustics, as well as the fields involved in the science of acoustics [55]. Hereby, the inner ring represents the underlying physical principles of sound creation and transmission. The middle ring illustrates the subdivisions of acoustics in terms of scientific categories. The outmost ring covers the technical and artistic fields that investigate and utilize acoustic signals. Acoustic Sensing for Medical applications covers aspects from Sonic Engineering and Vibration Analysis at the intersection of Electrical Engineering, Mechanical Engineering, and Computer Science to Bioacoustics and Medical Sciences.

Sound is a wave phenomenon, i.e. an oscillatory disturbance moving away from a source without transporting a significant amount of matter over a propagated distance through compressible media [83]. An acoustic wave can be received by the human ear or another form of detecting device such as a microphone or other vibration measurement device. The most important parameters to describe an acoustic wave are amplitude, the highest derivation of a wave from its central position, and frequency, the rate of complete oscillation cycles per second. Frequency f can be derived from the wavelength of a signal and is defined as:

$$f = \frac{c}{\lambda} \quad (2.1)$$

where λ is the wavelength and c is the speed of sound, which is dependent on the medium, with a value of $c_{air} = 343 \frac{m}{s}$ in air and $c_{water} = 1480 \frac{m}{s}$ in water. The unit of frequency is Hertz [Hz], where $1 \text{ Hz} = 1 \text{ m s}^{-1}$. An increasing density of a medium results in more closely packed molecules and therefore an increased speed of sound propagation. Figure 2.2 a) illustrates the relationship between frequency and wavelength of a sound wave. While an acoustic signal with only one frequency signal component like shown in Figure 2.2 a) can be constructed artificially, e.g. using a synthesizer, real world audio is usually characterized by

¹Definition retrieved from <https://www.merriam-webster.com/dictionary/acoustics> on September 23rd, 2022

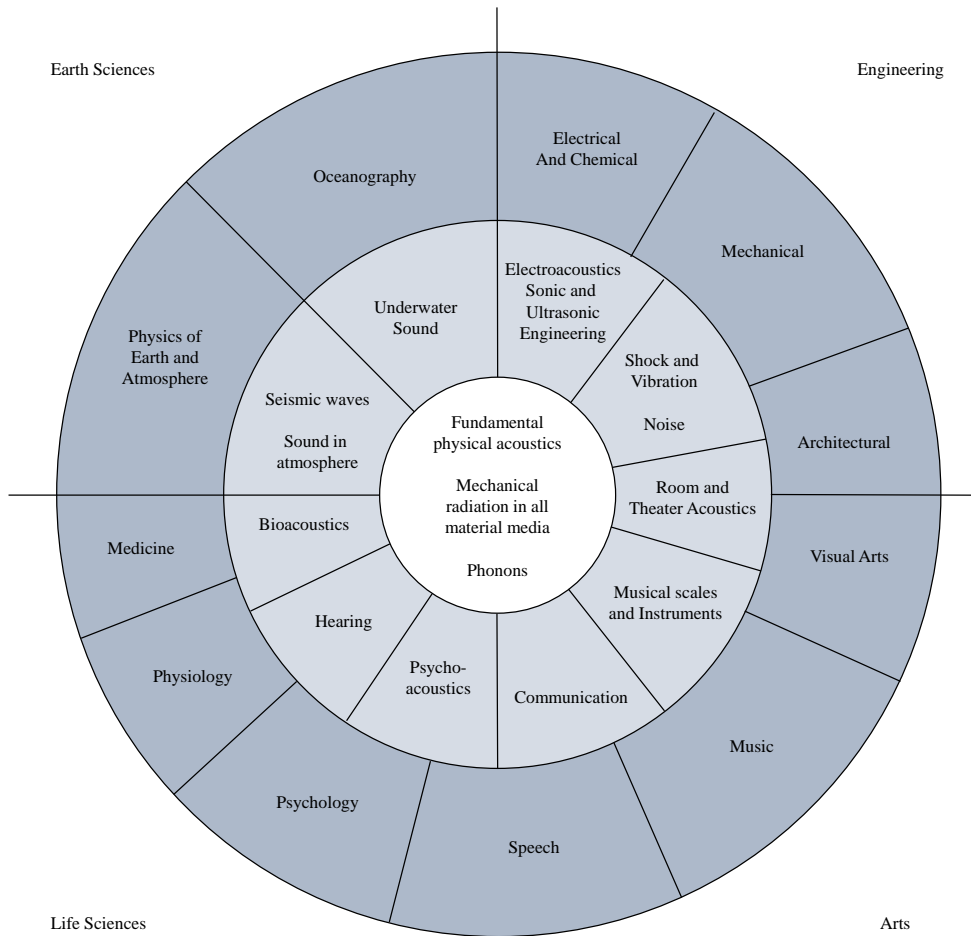


Fig. 2.1. The wheel of acoustics as defined by R. B. Lindsay, Adapted from [55].

the additive overlay of many different acoustic waves with different frequencies and result in a waveform as shown in Figure 2.2 b), an audio sample taken from the dataset created within the contribution described in section 3.2.1.

The audible range of the human hearing, which is also the range that most common microphones capture, is 20 Hz - 20 kHz. In the context of the present thesis, this audible range of acoustic waves is used to design acoustic sensing systems for medical interventions. The audible range of acoustic waves contains rich information and besides the benefit of interpretability by human listeners, working with the audible range of acoustic signals also has the advantage of being able to use professional high-quality microphones for capturing. Hereby, one should differentiate between two types of sound propagation types, air-borne and structure-borne sound propagation.

While the mechanism of sound propagation is identical, the capturing of air-and structure borne acoustic signals is associated with different challenges. Air-borne acoustic signals are propagated through the air and can be captured with a variety of different microphone tech-

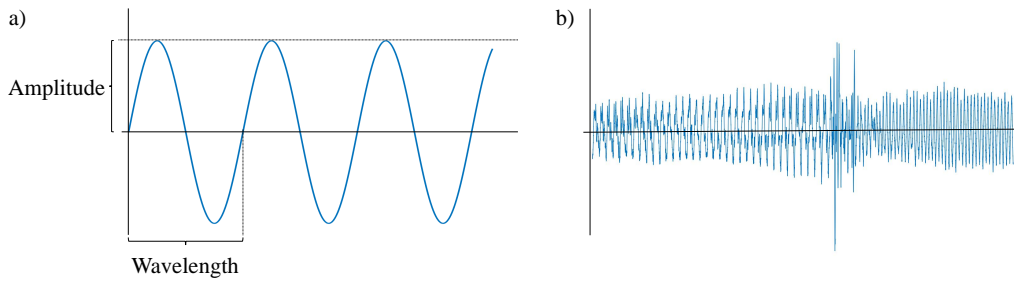


Fig. 2.2. a) A simple sine wave which illustrates the properties amplitude (highest deviation from central position) and wavelength (one full sine cycle). b) A real world waveform is composed of many different frequency components. The waveform is taken from the dataset created within the contribution presented in section 3.2.1 containing structure-borne drilling sound and a breakthrough event.

nologies, such as dynamic microphones (which use the same principle as speakers), condenser, or ribbon microphones. These types of microphones usually consist of a thin membrane which is excited by acoustic waves propagating through the air. This vibration of the membrane is then transformed to changes in electrical voltage. Air-borne microphones typically have different frequency responses and directionality patterns. Hereby, the directionality can range from spherical (captures sound in all directions) to cardioid (most sensitive to on-axis sounds) and directed (captures only sound in one direction and dampens sound from other directions).

Structure-borne microphones capture vibrations from structures and typically do not capture air-borne sounds. They use Micro-Electro-Mechanical Systems (MEMS) components such as accelerometers, or piezoelectric elements which transform structure-borne vibrations to changes in electrical voltage. In the context of the presented thesis, a custom highly sensitive modular contact microphone setup was developed which was used in the majority of research projects and is described in the contribution in section 3.2.1.

2.2 Processing Acoustic Signals

2.2.1 Audio Capturing and Preprocessing

While analog audio is still used, e.g. in music production, to capture the exact continuous signal with theoretically unlimited bandwidth, the prominent way of storing, streaming, and processing acoustic signals nowadays is digital audio. In order to digitize an analog voltage signal coming from a source, e.g. a contact microphone, it has to be amplified to a range required by the following Analog Digital Converter (ADC) stage. The input voltage range for ADCs is usually 0 V to 10 V or -5 V to 5 V. The standard method for digitizing audio signals is Pulse-Code Modulation (PCM) which consists of three stages, *Sampling*, *Quantization* and *Encoding*. In the *Sampling* stage, an electronic switch samples the analog signal in defined intervals, the so-called sampling frequency. The sampling frequency must be at least two times larger than the highest measured frequency to be able to recover the original analog signals from the discrete digital values, a correlation defined by the *Nyquist-Shannon sampling theorem* [102] as described in equation 2.2. Therefore, a common sampling frequency (e.g.

used for Compact Discs) for digital audio is 44 100 Hz, theoretically capturing the audible range up to 22 050 Hz.

$$f_{sample} \geq 2 * f_{signal} \quad (2.2)$$

The following *Quantization* stage assigns a digital value corresponding to the instantaneous value of the analog signal at each discrete sampling interval corresponding to the bit depth of the digital signal, e.g. 16, 24, or 32 bit. Finally, the *Encoding* stage converts the quantized sample to the binary PCM word code, the final digital uncompressed format of digital audio.

2.2.2 Audio Features

Feature extraction is the process of computing numerical features that capture significant information about the data while reducing the complexity. Commonly, a sliding window is applied to the original signal to slice it into smaller segments for the computation of features per individual time step. Hereby, often overlapping windows are used as the basis feature extraction to increase the amount of training data for learning-based algorithms.

For the analysis of acoustic signals (and signal processing in general), the frequency domain is a powerful representation because it contains the frequency components and their amplitudes present in the signal. Figure 2.3 visualizes a signal which is composed of two sine waves and the respective frequency spectrum which shows the two frequencies of the signal components as individual peaks. The transformation from signal into the spectral domain can be performed using the *Fourier Transform*. The discrete version of the Fourier Transform, the Discrete Fourier Transform (DFT) is described by the following formula:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-2\pi i k n / N}, \quad k = 0, \dots, N - 1 \quad (2.3)$$

where N is the number of samples per signal segment, $x(n)$ is a data point in the original signal, and $X(k)$ are the resulting Fourier coefficients. The DFT is commonly computed using Fast Fourier Transform (FFT) and its most prominent implementation, the *Cooley-Tukey* algorithm [17].

The following paragraphs present a small selection of common time and frequency domain audio features for illustration purposes and without claim of completeness. Examples for common audio features are *Root Mean Square (RMS) Energy*, a measure for the loudness of a specific audio segment which is calculated using the following formula:

$$RMS_t = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x(n)^2} \quad (2.4)$$

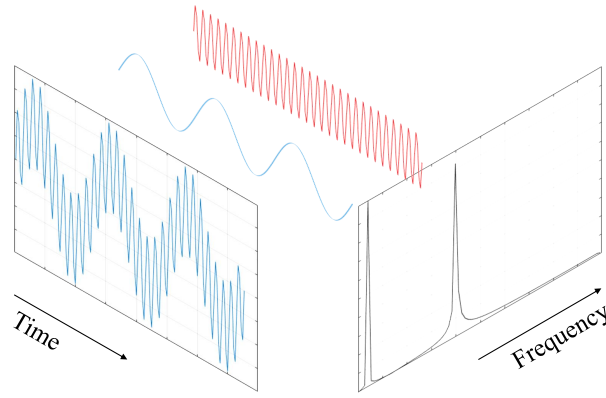


Fig. 2.3. The relationship between signal and frequency domain. The signal in the time domain is composed of two frequency components which are visible in the frequency spectrum as two peaks. The transformation from time to frequency domain is computed using the Fast Fourier Transform.

Another useful time domain feature is the *Zero Crossing Rate* of a signal segment which is computed using the equation 2.5, where *sgn* is the sign function. Zero Crossing Rate has been applied for speech and music recognition algorithms, especially for percussive sounds.

$$ZCR_t = \frac{1}{2} \sum_{n=0}^{N-1} |sgn(x(n)) - sgn(x(n+1))| \quad (2.5)$$

A common frequency domain feature is the *Spectral Centroid* which represents the center of mass of a frequency spectrum. It is associated to the brightness of a given sound segment and computed as the weighted mean of the signal's frequency components $x(k)$ using their magnitudes $f(k)$ as weights. K corresponds to the number of frequency bins in the spectrum.

$$SC_f = \frac{\sum_{k=0}^{K-1} f(k)x(k)}{\sum_{k=0}^{K-1} x(k)} \quad (2.6)$$

Perceptual studies indicated that humans perceive the pitch of an acoustic signal on a non-linear scale. Therefore, in 1937, Stevens et al. [107] proposed the *Mel scale*, a nonlinear mapping from frequency components which resembles the human perception of acoustic signals and can be computed using equation 2.7. For automated speech and music recognition, a widely used feature which applies the mel scale are Mel Frequency Cepstral Coefficients (MFCCs) which were introduced by Davis et al. in 1980 [21].

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (2.7)$$

MFCCs provide a compact representation of the frequency spectrum and are computed using a series of transformations applied to the power frequency spectrum derived from a time-domain signal. In the first step of the MFCC computation, a sliding window is applied to the signal

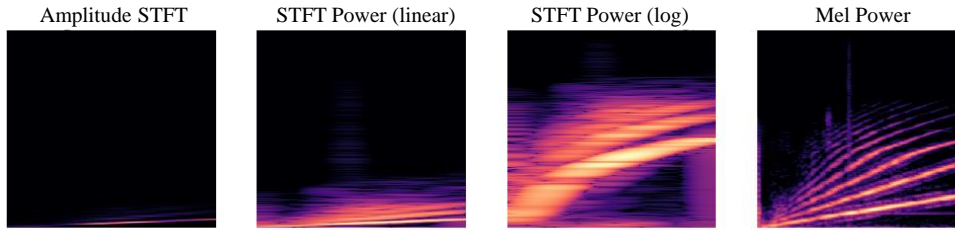


Fig. 2.4. Spectrogram visualization of an audio sample taken from the dataset of the work presented in section 3.2.2, all spectrograms are computed from the same source waveform. The visualizations illustrate, from left to right, an amplitude spectrogram with linear frequency scaling as computed using equation 2.8, a power STFT spectrogram as described in equation 2.9 and linear frequency scaling, the same power STFT spectrogram with a logarithmic frequency scaling, and rightmost, a Mel power spectrogram.

to slice it into (overlapping) windows and every window is Fourier-transformed. The Mel filterbank, a number of triangular filters spaced evenly on the Mel scale, is applied to the absolute power spectrum of a signal and the energy in each filter is summed up. In the last step, the Direct Cosine Transform (DCT) is computed from the summed up frequency bins which results in the desired *cepstrum*.

2.2.3 Spectrograms

Spectrograms are two-dimensional matrices, containing time on the x-axis and frequency components (bins) on the y-axis. Spectrograms as a representation for digital audio have the advantage of being a compact, grid-like data format while capturing the relevant characteristics from the frequency and time domain of the signal for further processing. However, it needs to be stated that the transformation from raw audio to a spectrogram is a lossy process and the reconstruction of waveforms from spectrograms is not trivial. The development of algorithms for performing this reconstruction is an active field of research which produced a variety of different methods such as the *Griffin-Lim* algorithm [33] or more recently also learning-based methods [110] that achieve superior reconstruction results.

The basic form of the spectrogram is the *Short Time Fourier Transform (STFT)* spectrogram. To compute the STFT, the data to be transformed is split into overlapping frames $x[n]$, each frame is Fourier transformed and the result is stored in a matrix. The formula to compute the STFT from a time-domain signal is defined as:

$$X(m, k) = \sum_{n=0}^{N-1} x(n + mH)w(n)e^{-2\pi i kn/N} \quad (2.8)$$

where X is a matrix containing the k^{th} Fourier coefficient for the m^{th} time frame as row and column values, respectively. The term $w(n)$ corresponds to a windowing function which smooths the start and end of the signal segment to avoid *spectral leakage* during FFT computation [59]. The variable H corresponds to the hop length used for STFT computation. As the amplitude of acoustic signals is commonly measured in decibels (dB), the STFT spectrum can be converted to a decibel scale using equation 2.9. If most of the information of the signal is

contained in the low frequency range, a logarithmic scale can be applied to the frequency axis of the spectrogram.

$$X_{dB}(m, k) = 10 \log_{10}(X(m, k)^2) \quad (2.9)$$

To transform a STFT spectrogram into a Mel spectrogram, the resulting spectrum of the Fourier transformation is filtered with triangular filters spaced evenly on the Mel scale as described in equation 2.7. The result of the different methods for spectrogram computation are illustrated and visualized in figure 2.4. Mel spectrograms have become the dominant feature representation for the design of audio deep learning systems and provide a compact data representation which achieves state-of-the-art results while having lower computational and data requirements compared to raw waveform input [87].

Furthermore, there are other spectrogram variants, such as Constant-Q Spectrograms which use an alternative transform, the Variable-Q (VQT) and Constant-Q Transform (CQT), to compute the frequency spectrum from a time-domain signal or Chromagrams which provide harmonic information which has proven to be useful for music related problems.

Data normalization is commonly used to normalize the numeric range of all data points in a dataset to improve the convergence of learning-based algorithms. Hereby, a common practice is to normalize all data points in a dataset to have a mean value of zero and a standard deviation of one, using the following formula where μ is the mean and σ is the standard deviation computed over the entire dataset:

$$x_{norm} = \frac{x - \mu}{\sigma} \quad (2.10)$$

Machine Learning-based Acoustic Sensing

3.1 Classical Approaches

Traditional machine learning methods for acoustic signal processing have been around for decades and are usually characterized by first extracting a number of features from the signal to build a feature vector which serves as the input to a learnable model. For the application to acoustic signals, the feature vector is usually composed of different time and frequency domain features of the source signal as described in section 2.2.1. The contained features are engineered by the expert designing the system in order to capture important characteristics and information about the source samples.

In the literature, a wide variety of classical machine learning methods using classical audio features as described in section 2.2.2 have been applied to classify and cluster acoustic signals. Examples for classical machine learning methods that have been applied to features computed from audio signals include Support Vector Machines (SVMs) [22], Random Forests (RF) [94], K-Nearest Neighbor (KNN) [58], Gaussian Mixture Models (GMMs) [41], or Hidden Markov Models (HMMs) [57]. Also in the medical domain, these methods have been applied to various acoustic signal-based applications, such as diagnostic applications, e.g. for the assessment of cartilage degeneration [7], and the analysis of interventional sounds of surgical drilling [123], milling [20], or needle insertion [38], as described in section 1.3.1 and 1.3.2.

The advances in computational resources and deep learning algorithms enabled the automatic feature extraction from more complex data representations than the reduced dimensionality of engineered and precomputed feature vectors, e.g. using Convolutional Neural Networks, and have outperformed classical machine learning techniques, which will be discussed in the following sections.

3.2 Spectrogram-based Audio Processing

For the design of learning-based acoustic sensing systems, automated feature extraction from digital audio signals using deep neural networks has outperformed classical handcrafted features in a variety of audio-related tasks [87]. Hereby spectrogram representations have advantages over feature extraction from raw waveforms in terms of computational resources and data requirements, as the filters are predetermined and do not additionally have to be implicitly learned by the model during the training process. Hereby, the mel spectrogram, which resembles the way humans perceive sound, has been shown to be a particularly

beneficial representation for deep learning-based audio processing tasks. While spectrogram-based deep learning methods have been applied to acoustic signal processing outside of the medical domain for various applications [87], these methods have only been scarcely used for medical applications so far.

In the medical domain, a frame-based acoustic sensing system has the advantages of capturing temporal aspects of the acoustic signal which can be controlled through the window size of the computed spectrogram, while being fast and responsive. In the following sections, we present two scientific contributions, the first work proposes a fast and reliable system for the detection of breakthrough events in orthopedic drilling for surgical error prevention. The second work presents a novel paradigm and methodology to detect loose pedicle screws as a complication after spinal fusion surgery based on vibroacoustic excitation and sensing.

3.2.1 Contribution (Interventional): Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery (Scientific Reports 2021)

Summary: The work illustrates the potential of using structure-borne acoustic signals for interventional surgical error prevention. In surgical interventions, human errors are inevitable and can result in severe patient harm and trauma. Previous studies have shown that acoustic signals are used by medical experts implicitly to guide them through the drilling process and have therefore great potential for the design of automated interventional guidance systems. We designed a system for the automated detection of drill breakthrough events, one of the most common tasks in orthopedic surgery, based on a custom modular contact microphone capturing high quality drilling vibration signals, placed directly on the patient's skin and a breakthrough detection pipeline based on mel spectrograms and a modified ResNet-18 classifier. Because the drill breakthrough events are short and occur scarcely compared to cortical bone drilling, the generated dataset is highly imbalanced which is explicitly handled in the training process utilizing the Focal Loss function. We compare the execution speed of different configurations of the pipeline which shows promising performance while being multiple times faster than the human reaction time. The method is validated in a human cadaveric experiment with six cadaveric hip specimens and evaluated using a cross validation scheme. The work advances the state-of-the-art in drill breakthrough detection by proposing the first deep learning-based detection pipeline which outperforms the results of previous studies. We evaluated the system in a realistic cadaveric setup in comparison to previous work which utilized artificial bone models or animal bone specimens for validation. Hence, the paper represents an important step towards the translation of acoustic-based drill breakthrough detection into clinical applications.

Contributions: The author of this thesis was responsible for formulating the problem and approach with medical consultations from Steven Maurer, Armando Hoch, Patrick Zingg, and Mazda Farshad, as well as for capturing the data in the experimental setup, designing the data processing, detection, and evaluation pipeline and writing the manuscript. All human cadaveric experiments were conceptualized and performed by the author, Steven Maurer, and

Armando Hoch, Nassir Navab and Philipp Fürnstahl contributed in the form of discussions and feedback throughout the whole project and for proofreading of the manuscript.

Copyright Statement: The article was published by Springer Nature in Scientific Reports and is licensed under a Creative Commons Attribution 4.0 International License.



OPEN

Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery

Matthias Seibold^{1,2✉}, Steven Maurer³, Armando Hoch³, Patrick Zingg³, Mazda Farshad³, Nassir Navab¹ & Philipp Fürnstahl^{2,3}

In this work, we developed and validated a computer method capable of robustly detecting drill breakthrough events and show the potential of deep learning-based acoustic sensing for surgical error prevention. Bone drilling is an essential part of orthopedic surgery and has a high risk of injuring vital structures when over-drilling into adjacent soft tissue. We acquired a dataset consisting of structure-borne audio recordings of drill breakthrough sequences with custom piezo contact microphones in an experimental setup using six human cadaveric hip specimens. In the following step, we developed a deep learning-based method for the automated detection of drill breakthrough events in a fast and accurate fashion. We evaluated the proposed network regarding breakthrough detection sensitivity and latency. The best performing variant yields a sensitivity of $93.64 \pm 2.42\%$ for drill breakthrough detection in a total execution time of 139.29ms. The validation and performance evaluation of our solution demonstrates promising results for surgical error prevention by automated acoustic-based drill breakthrough detection in a realistic experiment while being multiple times faster than a surgeon's reaction time. Furthermore, our proposed method represents an important step for the translation of acoustic-based breakthrough detection towards surgical use.

Surgical interventions are conducted by trained and experienced experts, however, human errors are inevitable. In the operating room, surgical errors can lead to significant and severe consequences for the patient, in the worst case to death¹. Prior studies showed that surgical factors account for more than 70% of intraoperative complications^{2,3}. For example in orthopedic surgery, iatrogenic femoral arterial⁴ and nerve⁵ injury are frequently happening complications caused by surgical errors. Detecting and preventing these incidents is crucial to improve the patient safety and the outcome of surgery⁶.

There is a variety of causes for surgical errors and resulting iatrogenic injuries. They range from anatomical differences between patients and proximity of risk structures⁷ or lack of surgical access and overview, for example in obese patients⁸, to pathologically altered tissue substance, e.g. in patients with osteoporosis⁹. Furthermore, the condition of the surgeon and the surgical staff plays an important role on the performance and therefore the outcome of the surgery, as lack of concentration and technical incapacity can lead to an increased risk of iatrogenic injury¹⁰.

To assess the patient-specific risk of treatment complications in orthopedic surgery, commonly a pre-operative plan based on the patient anatomy and medical imaging data, such as radiographs, computed tomography (CT) or magnetic resonance imaging (MRI)¹¹, is made. Furthermore, specialized imaging modalities, for example angiography⁷ or ultrasound^{12,13} are utilized to visualize anatomical risk structures such as nerves and arteries. Conventional navigation systems provide a way to transition pre-operative information into surgery by displaying it in relation to intraoperative information on an external monitor¹⁴ or even actively guide the surgeon through robotic assistance¹⁵, but prompt a need for additional optical tracking systems and time-consuming registration procedures which can introduce additional errors by registration failures¹⁶. Learning-based systems have great potential to support the surgeon during the intervention and enable augmented decision making based on real-time sensor data and additional learned knowledge¹⁷⁻¹⁹. They can be employed for active error prevention by detecting surgical states and important or adverse events during surgery in an automated fashion.

A relevant target task for an error prevention system in surgery is drill breakthrough detection. Drill breakthrough is defined by the drill perforating the bone and over-drilling beyond the far cortex into the adjacent soft

¹Computer Aided Medical Procedures (CAMP), Technical University of Munich, 85748 Munich, Germany. ²Research in Orthopedic Computer Science (ROCS), University Hospital Balgrist, University of Zurich, Balgrist Campus, 8008 Zurich, Switzerland. ³Balgrist University Hospital, 8008 Zurich, Switzerland. ✉email: matthias.seibold@tum.de

tissue. With the rise of machine learning methods, learning-based techniques also have been applied for the task of automated drill state and breakthrough detection using real-time sensor data and achieved promising results. Bone drilling is an essential part of orthopedic surgery and is conducted in about 95% of the interventions, for example to fixate bones with plates, external fixators and traction equipment²⁰. One of the most common ways of drilling in orthopedic procedures is to use free-hand power drills to manually pre-drill holes for bone screws. A study²¹ investigated free-hand drilling with a total number of 153 participating surgeons and found the average penetration of the soft tissue beyond the far cortex to be 6.31 mm which implies a great risk, especially when nerves, vessels or other vital structures are situated in close proximity to the target anatomy. The most important factor to stop the hand-operated drill as soon as possible after a breakthrough is the human reaction time. Even though trained surgeons have comparatively fast reaction times, their mean response time was measured to be in the range of 313 to 358ms which additionally decreases with advancing age²². A low-latency and robust detection system could enable a fast and automated stopping of the drill as soon as a breakthrough event is detected.

Drilling into a bone creates distinct vibrations resulting in the generation of acoustic signals which can be exploited for drill state and breakthrough analysis and have benefits over force/torque or current measurement approaches, such as easy integration and general applicability. Praamsma *et al.* showed in a study that experienced surgeons benefit from these audible sounds by utilizing them to support the drilling process²³. In this work, we used a custom piezo contact microphone to capture drill vibration signals in an experimental setup and propose a fast and robust deep-learning based drill breakthrough event detection method. The key contributions of our work are:

- We developed a custom high-sensitive piezo-based contact microphone prototype and impedance matching / pre-amplification stage for capturing structure-borne drill vibration signals non-invasively from the skin surface.
- We propose a low-latency and robust deep learning-based drill breakthrough detection method based on a modified ResNet-18 architecture, handling imbalanced data through the application of the *Focal Loss* function.
- We trained and validated our method on a dataset captured in an experimental setup using 6 unprepared human cadaveric hips including soft tissue.
- The proposed method outperforms the results of prior studies (using artificial bone models or prepared animal bone specimens) in a realistic cadaveric experiment.

State-of-the-art in acoustic-based drill breakthrough and drill state detection. Acoustic signals have been analyzed in prior work in order to detect both, drill breakthrough (penetration from bone into soft tissue) and drill state (type of tissue being drilled), using synthetic and animal bone models. The following paragraph gives an overview of the state-of-the-art in acoustic-based drill breakthrough and drill state detection and the transition from signal processing approaches to learning-based solutions in recent years.

One possible approach to implement automated drill state and breakthrough detection is based on force and torque measurements. Force/torque sensors offer reliable and accurate measurements of the force between drill bit and tissue and are therefore well suited for drill state and breakthrough detection^{24–26}. Recently, Torun *et al.* proposed a closed-loop method based on force sensor data to detect breakthrough events in an experimental setup operating on a sheep femur²⁷. As explained in their follow-up work²⁸, this approach has disadvantages for the application in real surgery, because they require physical modifications to the surgical device, in form of sensors attached to the drill which are costly and bulky. Another approach has been proposed to detect breakthrough events in electric drills by measuring changes in the current flow through the motor²⁹, which is however not suited for all types of surgical drills, such as pneumatic drills.

Because bone tissue consists of substructures with different density (cortical bone, cancellous bone and bone marrow), the friction between drill bit and tissue results in force and torque differences³⁰ and therefore in distinct vibrations for different tissue types during bone drilling. Drill breakthrough events result in an abrupt vibration change when perforating from high density cortical bone into soft tissue surrounding the bone. These distinct vibrations can be measured as acoustic signals. Therefore, acoustic-based drill state and breakthrough detection have been proposed in the literature as a low-cost and easy-to-integrate alternative to force/torque-based solutions.

Acoustic-based drill state detection has been introduced to classify different types of bone tissue during drilling by analyzing audio signals recorded from the area of operation. The first approaches achieved this task using signal processing-based techniques. A power spectral density based classification system was introduced by Sun *et al.* and evaluated in an experimental setup with five porcine scapulae using an air-borne room microphone³¹. Yu *et al.* proposed a sound-based solution for distinguishing between cortical and cancellous bone during surgical milling utilizing a wavelet package transform energy based state identification³².

Furthermore, learning-based approaches have been proposed to classify drill vibration signals based on prior knowledge. Boesnach *et al.* developed a method to analyze drill sounds in spine surgery by applying neural networks, support vector machines (SVM) and Hidden Markov Models (HMM)³³ to spectral density estimates. Zakeri *et al.* developed an experimental setup and learning-based classification method to distinguish between cortical and cancellous bone using six bovine tibiae by analyzing air-borne acoustic signals captured with a microphone³⁴. Their method is based on short-time Fourier transform (STFT) features in combination with a SVM classifier which achieved accuracies of up to 83%. In their follow-up research, they investigated different logistic regression, SVM, random forest (RF) and HMM classifiers and compared time and frequency features in regard to classification performance. The highest average accuracy of 84.3% could be achieved by using wavelet packet transform features³⁵. For the application in pedicle screw placement, a state recognition approach with

handcrafted features and a neural network classifier was proposed by Guan *et al.*³⁶. They showed that the detection of different bone layers using acoustic emission signals is more accurate and precise compared to force/torque measurements in a bovine test specimen. The recognition rate was reported as 84.2%.

The task of drill breakthrough detection differs from the drill state classification problem, as short breakthrough events have to be detected with high accuracy and as fast as possible. The aim is here to stop the drill after perforating the cortical bone to avoid damage to surrounding soft tissue. An automatic method to stop the drill when perforating a rat skull based on spectral density features and a SVM classifier was proposed by Pohl *et al.*³⁷ for the application in fully automated animal surgery. In a recent work, Torun *et al.* proposed a drill breakthrough detection method based on parametric power spectral density estimation. By computing four frequency features and applying a neural network classifier they could reach a breakthrough detection accuracy of $92.37 \pm 1.09\%$ in a 311.2ms time frame, using an artificial bone model²⁸ and acoustic signals captured with an air-borne microphone.

Drill breakthrough detection and state classification based on acoustic signals has been shown to be a promising approach to supervise the surgical drilling process^{28,31,32,34–36}. In prior work, studies have been conducted in an experimental setup using artificial bones or resected animal bone specimens and air-borne microphones attached to the drill or placed in close proximity to the area of operation. This is a limitation for the application in real surgery as the operating room is a noisy environment and the anatomy is not directly accessible because of surrounding soft tissue. To the best knowledge of the authors all previous studies implemented classical machine learning approaches, such as HMM, SVM or simple neural network classifiers. Recent advantages in deep learning methods for acoustic event classification have been shown to yield superior performances compared to classical approaches³⁸. These approaches typically employ higher dimensional feature representations, such as spectrograms, which enable the deep network to learn the optimal features itself during the training process. Furthermore, typical window lengths in acoustic breakthrough detection of 300ms, such as used in²⁸, achieved promising results and have been applied for robotic drilling applications, where the robot is programmed with a slow feeding rate. However, they are not sufficient for free-hand drilling supervision, as a surgeon can react just as fast as the automated system²². To translate automated drill breakthrough detection into clinical use, adaptations to the hardware and data acquisition setup, as well as the development of a robust and fast classification method are crucial. The main goal of this approach is preventing surgical errors in form of over-drilling into the adjacent soft tissue, therefore increasing the safety of intervention and reducing patient risk.

Cadaver experiments

In the following sections we will describe the experimental setup in detail, including recording hardware, conduction of the experiment and data preparation. Subsequently, the breakthrough detection method is introduced which is trained and validated on the dataset acquired in cadaver experiments.

Low-cost contact microphone, impedance matching and amplification. Piezo-electric elements are made of crystalline material and produce small voltages when force or pressure is applied. This principle can, when amplified, be utilized to record vibrations as structure-borne sound from a surface by using piezo-electric elements as contact microphones. Structure-borne sounds have been shown to have great potential for analysis and information retrieval in medical applications³⁹. Due to the physical nature of piezo-electric elements, the output impedance of the contact microphone lies typically in the range of several M Ω . This results in an impedance mismatch with microphone or line inputs of recorders or mixers, which usually have an input impedance in the range of a few k Ω . The mismatch results in high-pass filtering and poor transmission of signal energy in the low frequency region. Because we are interested in capturing also low-frequency components of the structure-borne vibration signal for breakthrough detection, an impedance matching stage is necessary. Furthermore, a high common-mode rejection ratio (CMRR) is desired to minimize electromagnetic interference. We use a 48V phantom-powered impedance matching circuit designed by Alex Rice (circuit design available under: <https://www.zachpoff.com/resources/alex-rice-piezo-preamplifier/>) and released under a Creative Commons Share-Alike 3.0 license. This circuit combines impedance-matching with a shielded and balanced connection, which suits the needs of our application. As contact sensor, we utilize a standard piezo disk with a diameter of 27mm. The contact microphone, impedance matching stage and analog/digital conversion stage are modular and connected through rugged XLR connector cables to allow different connection lengths for easy use.

To reduce the noise of the contact microphone and influence of electromagnetic fields, we electromagnetically shielded the entire circuit from piezo-element to the analog/digital converter and connected it to the system's ground. Furthermore, before shielding, the piezo disk was covered in epoxy resin (WEICON GmbH & Co. KG, Münster, Germany) to make it rugged and avoid noise introduced by moving cables. Every cable connection (from piezo-element to impedance matching stage and from impedance matching stage to audio interface) is designed as balanced line to remove any electrical interference during signal transmission. This results in a highly sensitive and low-noise contact microphone which can be attached to the skin of patients to capture structure-borne signals. For amplification and analog/digital conversion we use the PreSonus Studio 68 (PreSonus Audio Electronics, Inc., Baton Rouge, LA, USA) audio interface and the Audio Stream Input/Output (ASIO) low-latency driver. The microphone amplifiers have a frequency response of 20 Hz–20 kHz with a tolerance of ± 0.15 dB.

Another advantage of our setup is its modular design, which allows the contact microphone to be used as a disposable surgical instrument. All components of the contact microphone are low-cost (< 10 USD) and therefore suited for single-use. The whole hardware setup can be built with an associated cost of about 300 USD. Figure 1 provides an overview of the recording chain which was used in the experiments described in the following section.

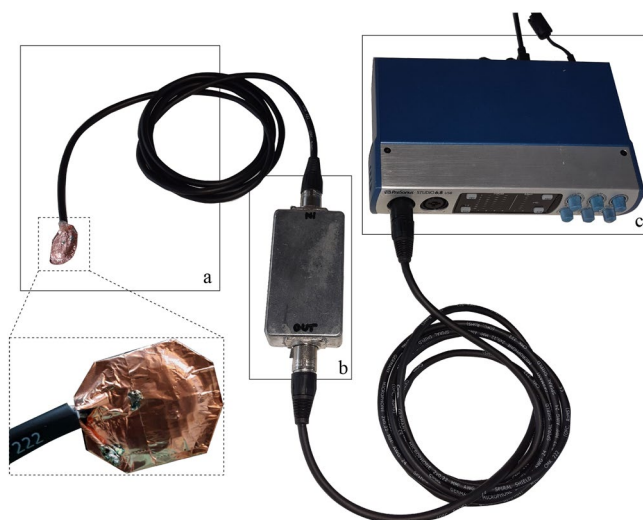


Figure 1. The recording chain consisting of (a) a shielded piezo contact microphone, (b) an impedance matching stage, and (c) an analog/digital conversion and amplification stage which allows to capture recordings from four sensors in parallel.

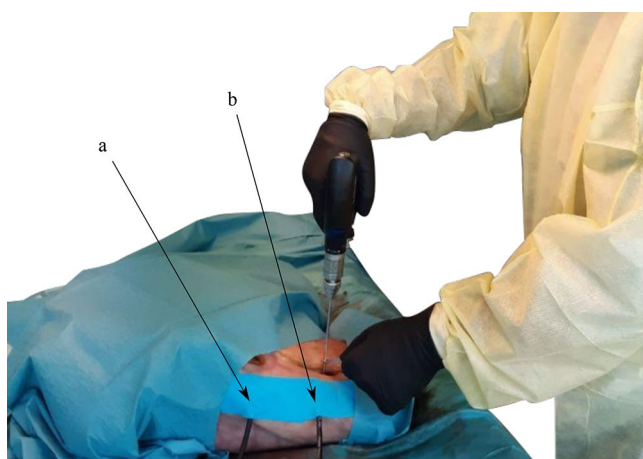


Figure 2. The experimental setup with a human cadaveric hip specimen. Two microphones are attached to the specimen's skin surface with kinesiology tape to permit synchronously recording in parallel. The contact sensors are placed (a) at the *trochanter major* and (b) next to the incision, in *diaphysis* position.

Cadaver study design and data acquisition. To build an experimental setup which is as realistic and close to a clinical scenario as possible, we used three fresh frozen and unresected human cadaveric hip specimens to generate a dataset for training and validating the proposed network. An ethical approval for all ex-vivo experiments (Kantonale Ethikkommission Zurich, protocol number: 2020-01913), as well as informed consent from all subjects involved in this study and/or their legal guardians has been obtained. The experiments were conducted by a trained physician according to relevant guidelines and regulations. None of the cadavers used in our experiment had a record of previously assessed osteoporosis. The specimens were thawed, prepared, and incision in the lower extremities were made to access the area of operation at the proximal femur until the upper shaft of the anterior femur. For the surgical approach we used a direct transmuscular access for optimal presentation of the femur. The incisions were executed by intersecting the midline of the *musculus quadriceps* longitudinally with a scalpel and detaching it anteriorly from the surface of the femur. We then attached two contact microphones to the specimen's skin surface using kinesiology tape. As illustrated in Fig. 2, one microphone was attached about 2cm next to the incision to minimize the distance that the acoustic waves propagate from source to microphone through the soft tissue, referred to as *diaphysis* position. The medical expert placed the second microphone on the skin where the *greater trochanter* is located. This placement was chosen because acoustic waves propagate well through bony tissue, a principle which has already been applied for bone quality assessment of long bones⁴⁰, and the bone structure is easily identifiable for consistent placement in a clinical scenario.

The effects of different sensor placement on drill breakthrough detection accuracy have been evaluated in this study as well and are presented in the section “[Comparison of microphone positioning](#)”.

We utilized a Colibri II battery powered drill (DePuy Synthes, Raynham, MA, USA) which is a standard power tool used in orthopedic surgery and a drill bit with a diameter of 3.2mm to drill holes into the femur. To create as realistic acoustic conditions as possible and to stabilize the drill on the periosteum, a tissue guard was used as seen in the Fig. 2. The drill bit was then placed in a right angle on the exposed periosteum of the anterior surface of the proximal femur and drilled in a continuous clockwise rotation. To be able to separate the recordings and assign them to the respective class, we recorded the breakthrough sequence from drilling through the second cortical layer of the femur until breaking through into the adjacent soft tissue. For each cadaver, we recorded data from both the left and right hip resulting in a total number of six individual bones. Overall, we captured audio recordings from 136 individual drill holes and respective breakthrough events in the experimental setup illustrated in Fig. 2. On average, about 22 holes were drilled in each femur, which corresponds to a realistic clinical scenario, as big Locking Compression Plates (LCP) plates with 20 and more holes exist for large bones.

After capturing, the recordings were manually cut, labelled and separated into two subsets $C := \{c_1, c_2\}$, where c_i denotes the respective class. We thoroughly identified each breakthrough sequence in the audio recordings by repeated acoustic and visual inspection in the respective spectrogram. In this context, the class c_1 , *cortical*, contains samples of drilling cortical bone and the class c_2 , *breakthrough*, contains samples of drill breakthrough events. All recordings processed within the digital audio workstation software REAPER. The samples were rendered without further application of software gain or processing. The recordings were captured with a sample rate of 44.1kHz and a bit depth of 24 bit. With a buffer size of 128 samples, the ASIO driver latency was measured as 6.8ms. In this configuration, up to four contact microphones can be recorded synchronously in parallel.

Breakthrough detection method

Pre-processing, feature extraction, data augmentation. Spectrogram features are the dominant representation in deep learning for audio signal processing³⁸. They have been shown to yield superior classification performances and achieve promising results in combination with convolutional neural network-based architectures for speech⁴², audio event detection⁴³, and medical applications⁴⁴. Log-mel spectrograms, a widely-used spectrogram variant, are two-dimensional matrices with time windows as columns, mel-bins (frequency) as rows, and amplitude as scalar values contained in the matrix. Because of their grid-like regular structure, they are well suited to be processed using CNN classifiers. To compute log-mel spectrograms, the discrete signal was first segmented with a rectangular sliding window into short frames $x : [0 : L - 1] := \{0, 1, \dots, L - 1\} \rightarrow \mathbb{R}$ of length L with 75% overlap. Short-Time Fourier Transformation (STFT) for each framed clip was computed using:

$$X(m, k) = \sum_{n=0}^{N-1} x(n + mH)w(n) \exp\left(\frac{-2\pi i kn}{N}\right), \quad \text{where} \quad (1)$$

$$w(n) = \frac{1}{2} \left[1 - \cos\left(2\pi \frac{n}{N-1}\right) \right], \quad n = 0, \dots, N-1 \quad (2)$$

Equation (2) denotes the *Hann* window function of length N , used for Eq. (1) to avoid *spectral leakage*⁴⁵. The sliding window was shifted across the signal, using a step size specified by the parameter H , in samples. The resulting matrix X is a STFT spectrogram and contains the k^{th} Fourier coefficient for the m^{th} time frame.

To evaluate the performance of the proposed system in regard to the window length used for spectrogram generation, we implemented different hop lengths $H = \{64, 32, 16\}$ for the window lengths evaluated in this paper, $L = \{4410, 2205, 1102\}$, respectively, to keep the final spectrogram dimensions constant. The result was converted to a power spectrogram representation by squaring the amplitude and subsequently mapped to a logarithmic decibel scale by computing:

$$X_{pow}(m, k) = 10 \log_{10}(X(m, k)_2) \quad (3)$$

For transferring the matrix to the Mel scale, the result was filtered in the spectral domain with a triangular shaped Mel filter bank. The triangular filters are spaced evenly on the Mel scale which can be calculated from frequency with Eq. (4).

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

The log-mel spectrogram representation provides sparse, high resolution features for audio sources⁴⁶. A total number of 256 Mel filter bands were applied to combine the Fast Fourier Transform (FFT) bins into Mel-frequency bins. All spectrograms were normalized by $X_{norm, mel} = (X_{mel} - \mu) / \sigma$, where (μ) is mean and (σ) is the standard deviation computed over the entire training data.

Because the length of the breakthrough events is in the range of 100 to 250ms and much shorter compared to non-breakthrough sequences in our dataset, the number of spectrograms computed for the non-breakthrough class is more than an order of magnitude larger. This results in a highly unbalanced dataset. To balance the dataset, we use a data augmentation strategy and apply it to the underrepresented class by varying the gain

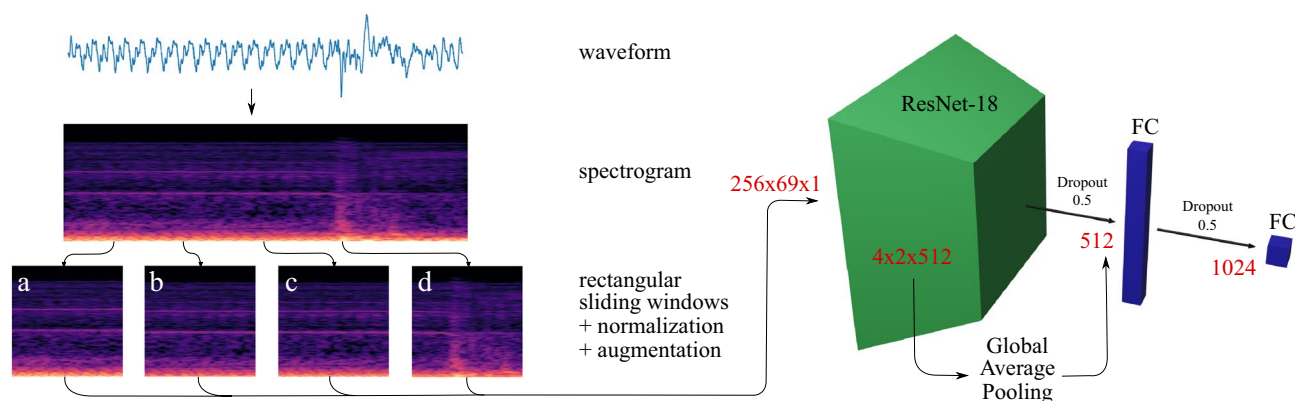


Figure 3. A breakthrough sequence with a total length of about 1 s, taken from the dataset acquired during the cadaver experiments. The raw waveform was split into frames by applying a rectangular sliding window and mel spectrogram features were computed. For better illustration, the spectrograms in this Figure are plotted using a colormap, however, the features used in the implementation of this work are two-dimensional only. Furthermore, the window length is chosen arbitrarily for better visualization and is not representative for the windows which have been evaluated in this work and are much shorter. Frames (a) to (c) correspond to the non-breakthrough class, in frame (d) the breakthrough event is present and visible in the spectrogram. The features were normalized and augmented which is described in detail in the section “Pre-processing, feature extraction, data augmentation”. A modified ResNet-18⁴¹ architecture, which is introduced in the section “Deep learning model and training”, was implemented to classify breakthrough events from spectrogram features. The output dimensions of each pipeline stage are given in red color.

(−5dB, 5dB), as well as applying time stretching (0.5, 0.7, 1.2, 1.5 times play rate) and pitch shifting (−3, −1, +1, +3 semitones) to the breakthrough event training samples.

A high-level overview of the pre-processing pipeline is illustrated in Fig. 3 in the left part of the Figure. All spectrograms were computed using the python library *librosa 0.7.2*⁴⁷, are of size 256x69x1 and serve as input for the convolutional neural network architecture which is described in detail in the following paragraph.

Deep learning model and training. The deep residual network (ResNet) architecture⁴¹ has been shown to perform exceptionally well on spectrogram-based audio classification tasks⁴⁸. Because our aim is to develop a low-latency and reactive system we chose to implement a 18-layer ResNet variant which enables fast inference⁴⁹. We found empirically to achieve the best results with a slightly modified architecture, stacking a global average pooling layer⁵⁰, a dropout layer with a dropout rate of 0.5, a fully connected (FC) layer with 1024 neurons, another dropout and an output FC layer on top of ResNet-18’s final batch normalization layer. By introducing additional dropout layers for regularization, we reduce the model’s tendency towards overfitting. The final model has a total number of 11,715,393 parameters and its architecture is illustrated in Fig. 3.

To handle the problem of imbalanced data, we apply the *Focal Loss*⁵¹ as loss function for training. For imbalanced datasets, standard crossentropy is inefficient, as most samples fed to the network are classified with large confidence and therefore contribute no useful learning signal. The *Focal Loss* influences the network to focus on the underrepresented class which in our case corresponds to the breakthrough events that are crucial to detect with high accuracy for our particular application. The *Focal Loss* function is defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad \text{where } p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \quad (5)$$

The factors γ and α_t are introduced as *focusing* and *balancing* parameter, respectively. In our implementation we use the by Lin *et al.* empirically determined optimal values $\alpha_t = 0.25$ and $\gamma = 2$ ⁵¹. The variable p_t is defined for convenience, where p corresponds to the estimated probability for the class with label $y = 1$. We trained the model end-to-end on the spectrogram features explained in the section “Pre-processing, feature extraction, data augmentation” using the *Adam* optimizer and reduced the learning rate when stagnating loss was observed over three epochs by a factor of 10.

Model, training and inference were implemented using the open-source deep learning library *TensorFlow 2.2* and run on a NVIDIA GeForce RTX 2080 SUPER GPU. All results presented in the following sections have been evaluated using 5-fold cross-validation.

Results

We split the evaluation section into three parts. First, we present the best performing variant of the proposed detection method and analyze the influence of design decisions on our detection pipeline. Afterwards, we compare the two microphone positions described in the section “Cadaver experiments” by analyzing the performance using synchronously acquired audio data. Subsequently, we evaluate different sliding window lengths L to assess the trade-off between detection latency and accuracy.

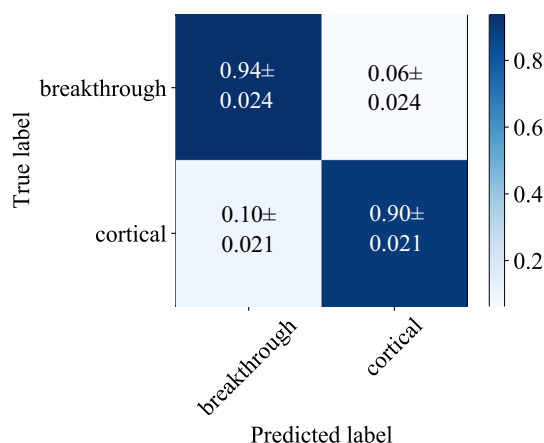


Figure 4. The normalized confusion matrix for a rectangular sliding window size of length $L = 4410$ samples which corresponds to a time frame of 100ms, recorded in the *greater trochanter* position.

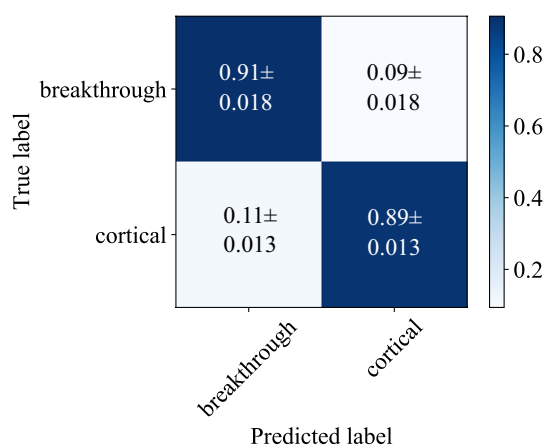


Figure 5. The normalized confusion matrix for a rectangular sliding window size of length $L = 4410$ samples which corresponds to a time frame of 100ms, recorded in the *diaphysis* position.

Detection accuracy and performance. Figure 4 shows the confusion matrix for the best performing variant of our proposed algorithm, evaluated on the independent test set with a 100ms rectangular sliding window and for the data recorded in *greater trochanter* position. We measured a mean overall accuracy of 97.29% in the training and 91.90% in the test phase. The recall (sensitivity) of correctly detecting breakthrough events, which is the main performance measure for our application, is measured as $93.64 \pm 2.42\%$.

Compared to the original ResNet-18 implementation⁴¹, we implemented several modifications which resulted in a performance gain. First, we modified the ResNet-18 architecture by including additional dropout and dense (FC) layers as described in the section “[Deep learning model and training](#)”. Through these modifications, we could boost the performance of the classification pipeline by about 3.5% for breakthrough recall. By implementing the *Focal Loss* as described in the section “[Deep learning model and training](#)” instead of the standard cross-entropy loss, we could furthermore increase the model’s sensitivity for breakthrough event detection by 11.2%.

Comparison of microphone positioning. We synchronously captured all recordings with two microphones in different positions as described in the section “[Cadaver experiments](#)”. One microphone was positioned directly above the *greater trochanter* to exploit the bone conductivity of acoustic waves. The second microphone was placed next to the incision (*diaphysis* position) to minimize the distance that the sound waves propagate through the soft tissue. The experimental setup and microphone positions are illustrated in Fig. 2. We treat the data acquired from the individual microphones as independent datasets to compare the microphone positions in regard to the resulting detection accuracy using a rectangular sliding window length of $L = 4410$ samples (100ms).

In comparison to the results for the data recorded in *greater trochanter* position and illustrated in Fig. 4, it can be observed in Fig. 5 that positioning the microphone in *diaphysis* position yields inferior detection performance. The recall for detecting breakthrough events is lowered by roughly 3% to $90.61 \pm 1.77\%$.

Window length (ms)	Sensitivity breakthrough %	Sensitivity cortical %
25	84.38 ± 2.69	75.58 ± 2.55
50	88.49 ± 3.88	82.52 ± 1.84
100	93.64 ± 2.42	90.16 ± 2.09

Table 1. Comparison of window length.

Pipeline stage	Execution (ms)	Execution (ms)	Execution (ms)
ASIO driver latency	6.8	6.8	6.8
Window length	25	50	100
Spectrogram generation	7.01	6.92	6.98
ResNet-18 inference	25.03	25.02	25.51
Total execution time	63.84	88.74	139.29

Table 2. Pipeline execution times.

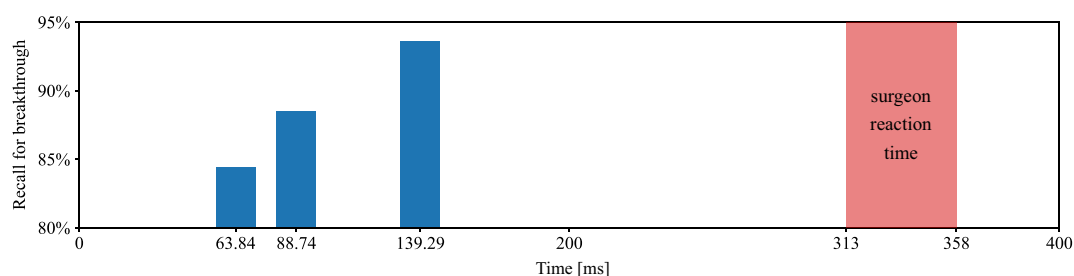


Figure 6. Pipeline execution speed and detection performance in comparison with surgeon reaction time, which has been measured to be in the range of 313 to 358ms and degrading with advancing age²². The execution times 63.84ms, 88.74ms and 139.29ms correspond to window lengths of 25ms, 50ms and 100ms, respectively, as explained in the section “Comparison of window lengths”.

For evaluating the influence of the window length L in the following section, we focus on the dataset recorded by the microphone in the *greater trochanter* position.

Comparison of window lengths. The length L of the rectangular sliding window determines the detection latency. With decreasing window length, the system is able to provide a detection result faster, as the audio chunk has to be acquired before it can be fed into the classification pipeline. However, the smaller the audio frame, the less information can be used by the network for feature extraction. We evaluated three window lengths, 100ms, 50ms and 25ms to gain insights about the performance of the proposed pipeline in comparison to the latency. We did not evaluate larger window lengths, as the shortest samples of breakthrough events are only a few ms longer than 100ms.

To investigate the influence of shorter window lengths, we systematically reduced the window length and evaluated the detection performance which is illustrated in Table 1. By lowering the frame length, the sensitivity for breakthrough detection and for classification of non-breakthrough samples is reduced. In general, it can be observed that the model’s performance decreases with shorter window lengths.

In Table 2, we show the measured execution time for each part of the proposed pipeline and the total execution time for one pass for a sample through the pipeline, given for different window lengths. All presented results have been averaged over 100 passes through the pipeline. Because we keep the spectrogram dimensions constant, spectrogram generation and ResNet-18 inference show very similar measured duration. We compare the above presented results to the average surgeon reaction time as measured in a previous study by Boom-Saad *et al.*²² in Fig. 6.

Discussion

Automated drill breakthrough detection is a promising approach for reducing the risk of surgical errors and iatrogenic injuries during the drilling process. To the best knowledge of the authors, our proposed method outperforms the results of all previous published work, for example Torun *et al.*, who used 300ms windows combined with handcrafted frequency features and achieved a detection sensitivity of $92.37 \pm 1.09\%$ for breakthrough detection in a simplified experimental setup based on a single artificial bone model²⁸. Our best performing algorithm variant achieves a breakthrough detection sensitivity of $93.64 \pm 2.42\%$ using a 100ms window. We

showed that our method performs well, even when reducing window lengths down to 25ms. We are not only using much shorter window lengths, but also transferred breakthrough detection to pre-clinical experiments using human cadavers with soft tissues to mimic a realistic surgical intervention.

Low-latency detection is crucial to stop the drill as fast as possible when a breakthrough event is observed. The total latency of the best performing variant of our algorithm amounts to 139.29ms. However, by reducing the window length we still achieve sensitivities of $88.49 \pm 3.88\%$ in 88.74ms and $84.38 \pm 2.69\%$ in 63.84ms for breakthrough detection. The larger the observed window, the more temporal context is provided to the model as basis for feature extraction and classification. As the performance of our solution still reaches fairly high classification accuracies with short window length, the trade-off between accuracy and latency has to be chosen for the particular application. The detection speed of our pipeline clearly outperforms the human reaction time and has therefore great potential to increase the safety during drilling in surgery.

Using easy-to-integrate contact microphones, we acquired structure-borne audio signals during drilling execution directly from the skin surface with very little noise disturbances. Acquiring a dataset in a realistic scenario such as cadaveric experiments has the advantage of capturing realistic characteristics of structure-borne acoustic signal dampening through soft tissue which is not possible to simulate with artificial bone models or prepared bones. Concerning the positioning of the contact microphones, our results show that the sounds captured in *greater trochanter* position yield better classification performance, compared to placing the microphone close to the incision in the *diaphysis* position. By exploiting the bone conductivity of acoustic waves and at the same time providing a reproducible positioning of the microphone, the *greater trochanter* position is optimal for the task of drill breakthrough detecting with contact sensors in hip surgery.

We believe that the deep-learning based analysis of structure-borne acoustic signals is a promising approach to supervise the surgical drilling process and that the proposed solution paves the path for deployment and testing the approach in real surgery. However, to translate the proposed system into clinical use in the operating room, the following limitations of the presented study have to be overcome. A clinical study is necessary to evaluate the reliability and robustness of the solution in-vivo. Furthermore, the performance of the algorithm could potentially be further improved by increasing the size of the dataset to expand the model's capability for generalization, including multiple surgeons and different anatomies. The proposed hardware setup, illustrated in detail in Fig. 1, is modular, low-cost and the contact microphones can be replaced easily. However, the sterilizability of the electrically shielded contact sensor (part *a* in Fig. 1) has to be investigated and validated. We did not measure the frequency transmission characteristics of the deployed circuitry explicitly, however the presented configuration enables high quality and low-noise audio recordings and increased bandwidth of the piezo element through impedance matching. Even though we thoroughly labelled each breakthrough sequence in the audio recordings by repeated acoustic (with professional studio-grade headphones) and visual inspection in the respective spectrogram (with high resolution in time), small uncertainties in the ground truth labelling process cannot be ruled out.

Currently, our system is running on a development computer, using high-end and high power hardware. To transfer the developed solution to an embedded solution, strategies such as model quantization can be employed to decrease the model size and resource requirements⁵². In addition, it is crucial to stop the drill as soon as a breakthrough event is detected to reliably increase the safety of surgical procedures by automated drill breakthrough detection. To this end, a stopping mechanism or circuitry has to be integrated into the drill which should be able to stop the drill with as minimal additional latency as possible.

Conclusion

In this paper, we present a deep-learning based approach for automated drill breakthrough detection in orthopedic interventions using acoustic emission signals. We developed a hardware setup employing piezo-based contact microphones to capture vibration signals non-invasively from the skin surface. The proposed experimental setup was utilized to capture a dataset of drill vibration signals from six human cadaveric hips.

Our classification pipeline reaches a sensitivity of $93.64 \pm 2.42\%$ on the task of drill breakthrough detection, in a total execution time of 139.29ms. Faster versions of our solution yield a sensitivity of $88.49 \pm 3.88\%$ in 88.74ms and $84.38 \pm 2.69\%$ in 63.84ms execution time. We show, that the proposed system is able to detect breakthrough events with high accuracy while being multiple times faster than the reaction time of trained surgeons. In addition, we evaluated different positioning of the contact sensors and observed that best results can be obtained by exploiting the conductivity of acoustic waves through bone tissue and placing the microphone as close as possible to subcutaneous bony structures.

The proposed solution has great potential to be used as a system for error prevention in surgery by preventing a damage to soft tissue and vital adjacent structures during bone drilling. Because drilling is an essential part in the vast majority of orthopedic interventions, the proposed system could have a great impact on patient safety and surgery outcome. Our exemplary application shows that acoustic sensing offers a very accurate, easy-to-integrate and low-cost approach to prevent errors in surgery which can be easily transferred to other surgical applications.

Received: 26 October 2020; Accepted: 3 February 2021

Published online: 17 February 2021

References

1. Sarker, S. K. & Charles, V. Errors in surgery. *Int. J. Surg.* **3**, 75–81 (2005).
2. Farshad, M., Bauer, D. E., Wechsler, C., Gerber, C. & Aichmair, A. Risk factors for perioperative morbidity in spine surgeries of different complexities: a multivariate analysis of 1,009 consecutive patients. *Spine J.* **18**, 1625–1631 (2018).
3. Farshad, M., Aichmair, A., Gerber, C. & Bauer, D. E. Classification of perioperative complications in spine surgery. *Spine J.* **20**, 730–736 (2020).

4. Giswold, M. E., Landry, G. J., Taylor, L. M. & Moneta, G. L. Iatrogenic arterial injury is an increasingly important cause of arterial trauma. In *90th Annual Meeting of the North Pacific Surgical Association* (Portland, Oregon, USA, 2003).
5. Moore, A. E. & Stringer, M. D. Iatrogenic femoral nerve injury: a systematic review. *Surg. Radiol. Anat.* **33**, 649–658 (2011).
6. Etchells, E., O'Neill, C. & Bernstein, M. Patient safety in surgery: error detection and prevention. *World J. Surg.* **27**, 936–941 (2003).
7. Ulm, A. J. *et al.* Normal anatomical variations of the v3 segment of the vertebral artery: surgical implications. *J. Neurosurg. Spine* **10**, 451–460 (2010).
8. Russo, M. W., Macdonell, J. R., Paulus, M. C., Keller, J. M. & Zawadzky, M. W. Increased complications in obese patients undergoing direct anterior total hip arthroplasty. *J. Arthroplasty* **30**, 1384–1387 (2015).
9. Mears, D. Surgical treatment of acetabular fractures in elderly patients with osteoporotic bone. *J. Am. Acad. Orthop. Surg.* **7**, 128–147 (1999).
10. Nurok, M., Czeisler, C. & Lehmann, L. Sleep deprivation, elective surgical procedures, and informed consent. *N. Engl. J. Med.* **363**, 2577–2579 (2010).
11. Hernandez, D., Garimella, R., Eloorai, A. E. M. & Daniels, A. H. Computer-assisted orthopaedic surgery. *Orthop. Surg.* **9**, 152–158 (2017).
12. Attinger, C. E., Meyr, A. J., Fitzgerald, S. & Steinberg, J. S. Preoperative doppler assessment for transmetatarsal amputation. *J. Foot Ankle Surg.* **49**, 101–105 (2010).
13. Ricci, S. Ultrasound observation of the sciatic nerve and its branches at the popliteal fossa: always visible, never seen. *Eur. J. Vasc. Endovasc. Surg.* **30**, 659–663 (2005).
14. Mavrogenis, A. F. *et al.* Computer-assisted navigation in orthopedic surgery. *Orthopedics* **36**, 631–642 (2013).
15. Lonner, J. H. & Moretti, V. M. The evolution of image-free robotic assistance in unicompartmental knee arthroplasty. *Am. J. Orthop.* **45**, 249–254 (2016).
16. Zhang, J. N., Fan, Y. & Hao, D. J. Risk factors for robot-assisted spinal pedicle screw malposition. *Sci. Rep.* **9**, 1–6 (2019).
17. Poduval, M., Ghose, A., Manchanda, S., Bagaria, V. & Sinha, A. Artificial intelligence and machine learning: a new disruptive force in orthopaedics. *Indian J. Orthop.* **54**, 109–122 (2020).
18. Navarrete-Welton, A. J. & Hashimoto, D. A. Current applications of artificial intelligence for intraoperative decision support in surgery. *Front. Med.* **15**, 369–381 (2020).
19. Hashimoto, D. A., Rosman, G., Rus, D. & Meireles, O. R. Artificial intelligence in surgery: promises and perils. *Ann. Surg.* **268**, 70–76 (2018).
20. Allotta, B., Giacalone, G. & Rinaldi, L. A hand-held drilling tool for orthopedic surgery. *IEEE/ASME Trans. Mechatron.* **2**, 218–229 (1997).
21. Clement, H., Heidari, N., Grechenig, W., Weinberg, A. M. & Pichler, W. Drilling, not a benign procedure: laboratory simulation of true drilling depth. *Injury* **43**, 950–952 (2012).
22. Boom-Saad, Z. *et al.* Surgeons outperform normative controls on neuropsychologic tests, but age-related decay of skills persists. *Am. J. Surg.* **95**, 205–209 (2008).
23. Praamsma, M. *et al.* Drilling sounds are used by surgeons and intermediate residents, but not novice orthopedic trainees, to guide drilling motions. *Can. J. Surg.* **51**, 442–446 (2008).
24. Hsu, Y.-L., Lee, S.-T. & Lin, H.-W. A modular mechatronic system for automatic bone drilling. *Biomed. Eng. Appl. Basis Commun.* **13**, 168–174 (2001).
25. Lee, W.-Y. & Shih, C.-L. Force control and breakthrough detection of a bone drilling system. In *2003 IEEE International Conference on Robotics and Automation* (Taipei, Taiwan, 2003).
26. Aziz, M. H., Ayub, M. A. & Jaafar, R. Real-time algorithm for detection of breakthrough bone drilling. *Procedia Eng.* **41**, 352–359 (2012).
27. Torun, Y. & Öztürk, A. A new breakthrough detection method for bone drilling in robotic orthopedic surgery with closed-loop control approach. *Ann. Biomed. Eng.* **48**, 1218–1229 (2020).
28. Torun, Y. & Pazarci, O. Parametric power spectral density estimation-based breakthrough detection for orthopedic bone drilling with acoustic emission signal analysis. *Acoust. Aust.* **48**, 221–231 (2020).
29. Puangmali, P., Jettumronglerd, S., Wongratanaphisan, T. & Cole, M. O. T. Sensorless stepwise breakthrough detection technique for safe surgical drilling of bone. *Mechatronics* **65**, 102306 (2020).
30. Wang, W., Shi, Y., Yang, N. & Yuan, X. Experimental analysis of drilling process in cortical bone. *Med. Eng. Phys.* **36**, 261–266 (2014).
31. Sun, Y. State. *et al.* State recognition of bone drilling with audio signal in robotic orthopedics surgery system. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3503–3508 (Chicago, Illinois, USA, 2014).
32. Yu, D., Yuan, X. & Jianxun, Z. State identification based on sound analysis during surgical milling process. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (Zhuhai, China, 2015).
33. Boesnach, I., Hahn, M., Moldenauer, J. & Beth, T. Analysis of drill sound in spine surgery. *Perspective in Image-Guided Surgery*, 77–84 (2004).
34. Zakeri, V. & Hodgson, A. J. Classifying hard and soft bone tissues using drilling sounds. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Seogwipo, South Korea, 2017).
35. Zakeri, V. & Hodgson, A. J. Automatic identification of hard and soft bone tissues by analyzing drilling sounds. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**, 404–414 (2019).
36. Guan, F. *et al.* State recognition of bone drilling based on acoustic emission in pedicle screw operation. *Sensors (Basel)* **18**, 1484 (2019).
37. Pohl, B. M., Jungmann, J. O., Christ, O. & Hofmann, U. G. Automated drill-stop by svm classified audible signals. In *34th Annual International Conference of the IEEE EMBS* (San Diego, California, USA, 2012).
38. Purwins, H. *et al.* Deep learning for audio signal processing. *IEEE J. Sel. Top. Sign. Proces.* **14**, 206–219 (2019).
39. Illanes, A. *et al.* Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing. *Sci. Rep.* **8**, 1–11 (2018).
40. Tatarinov, A., Sarvazyan, N. & Sarvazyan, A. Use of multiple acoustic wave modes for assessment of long bones: model study. *Ultrasonics* **43**, 672–680 (2005).
41. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
42. Stolar, M., Lech, M., Bolia, R. S. & Skinner, M. Acoustic characteristics of emotional speech using spectrogram image classification. In *International Conference on Signal Processing and Communication Systems (ICSPCS)* (2018).
43. Mesaros, A. *et al.* Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**, 379–393 (2018).
44. Ostler, D. *et al.* Acoustic signal analysis of instrument-tissue interaction for minimally invasive interventions. *Int. J. Comput. Assist. Radiol. Surg.* **15**, 771–779 (2020).
45. Lyon, D. A. The discrete Fourier transform, part 4: spectral leakage. *J. Object Technol.* **8**, 23–34 (2009).
46. Lewicki, M. S. Efficient coding of natural sounds. *Nat. Neurosci.* **5**, 356–363 (2002).
47. McFee, B. *et al.* librosa: audio and music signal analysis in python. In *14th Python in Science Conference*, 18–25 (2015).

48. Hershey, S. *et al.* Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 131–135 (2017).
49. Bianco, S., Cadene, R., Celona, L. & Napolitano, P. Benchmark analysis of representative deep neural network architectures. *IEEE Access* **6**, 64270–64277 (2018).
50. Lin, M., Chen, Q. & Yan, S. Network in network. arxiv:1312.4400v3 (2014).
51. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007 (2017).
52. Paschali, M., Gasperini, S., Roy, A. G., Fang, M. Y. & Navab, N. 3dq: compact quantized neural networks for volumetric whole brain segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13–17, 2019* (2019).

Acknowledgements

This work is part of the SURGENT project and was funded by University Medicine Zurich/Hochschulmedizin Zürich. Matthias Seibold and Nassir Navab are partly funded by the Balgrist Foundation in form of the guest professorship at Balgrist University Hospital.

Author contributions

M.S. proposed and implemented the method, wrote the main manuscript text and prepared all figures. M.S., S.M. and A.H. conducted the cadaveric experiments. P.Z. and M.F. supported the experiments from the clinical side and reviewed the manuscript. N.N. and P.F. supervised the work from the technical side and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

3.2.2 Contribution (Diagnostic): A new sensing paradigm for the vibroacoustic detection of pedicle screw loosening (Not peer-reviewed)

This work did not undergo peer-review and is not relevant for the grading of this dissertation.

Summary: Pedicle screw loosening is one of the most common complications after spinal fusion surgery which is characterized by the implant losing hold in the surrounding bone and causing severe pain for affected patients. The current state-of-the-art for the detection of loose pedicle screws is CT imaging which is not able to confirm implant loosening in a substantial number of cases. Therefore, there is a clinical need to develop novel paradigms and methodologies for the diagnosis of pedicle screw loosening. In this work, we propose a reliable, easy-to-integrate, radiation-free and non-destructive solution for the intraoperative detection of loose screws using an excitation device placed on top of the spinous process of the vertebra of interest and exciting the bone tissue using a sine sweep vibration. Contact microphones attached to the screw heads capture the resulting vibration characteristics transmitted through the excited bone to the implant. We design a custom pipeline based on a SE-ResNet-18 to detect a loose pedicle screw based on the vibration propagation pattern represented in mel spectrograms. We built an experimental setup consisting of four human cadaveric spine specimens and simulated the screw loosening using a 3D printed drill guide which was designed by a biomedical engineer according to the characteristics of screw loosening in patients based on the related literature. We validate the screw loosening simulation by analyzing the relative movement of vertebra and implant in a biomechanical testing machine using realistic loads during movement and the analysis of optical tracking data. Our results show promising performance for the development of an alternative assessment method for pedicle screw loosening. We furthermore believe that our work can be the starting point for the design of smart implants in spinal fusion surgery.

Contributions: The author of this thesis was responsible for formulating the problem and approach with medical consultations from Sandro Hodel, Mazda Farshad, and Christoph J. Laux, as well as for capturing the data in the experimental setup, designing the data processing, detection, and evaluation pipeline and writing the manuscript. All human cadaveric experiments were conceptualized and performed by the author and Christoph J. Laux. Bastian Sigrist supported the project by creating the preoperative plan based on 3D models segmented from CTs of the anatomy and designed the 3D printed drill guides. Tobias Götschi and Jonas Widmer supported the project for the design of the biomechanical experimental setup. Nassir Navab and Philipp Föhnstahl contributed in the form of discussions and feedback throughout the whole project and for proofreading of the manuscript.

Copyright Statement: This work is not peer-reviewed. The preprint has been published under <https://arxiv.org/abs/2210.16170>

A new sensing paradigm for the vibroacoustic detection of pedicle screw loosening

Matthias Seibold^{1,2}, Bastian Sigrist¹, Tobias Götschi³,
Jonas Widmer³, Sandro Hodel⁴, Mazda Farshad⁴,
Nassir Navab², Philipp Fürnstahl¹, Christoph J. Laux⁴

¹Research in Orthopedic Computer Science, Balgrist University Hospital, University of Zurich, Switzerland

²Computer Aided Medical Procedures, Technical University Munich, Germany

³Spine Biomechanics, Balgrist University Hospital, ETH Zurich, Switzerland

⁴Department of Orthopedics, Balgrist University Hospital, University of Zurich, Switzerland

Abstract

There is an unmet clinical need for developing novel methods to complement and replace the current radiation-emitting imaging-based methods for the detection of loose pedicle screws as a complication after spinal fusion surgery which fail to identify a substantial amount of loose implants. In this work, we propose a new methodology and paradigm for the radiation-free, non-destructive, and easy-to-integrate detection of pedicle screw loosening based on vibroacoustic sensing. Furthermore, we propose a novel simulation technique for pedicle screw loosening, which is biomechanically validated. For the detection of a loose implant, we excite the vertebra of interest with a sine sweep vibration at the spinous process and attach a custom highly-sensitive piezo contact microphone to the screw head to capture the propagated vibration characteristics which are analyzed using a detection pipeline based on spectrogram features and a SE-ResNet-18. To validate the proposed approach, we conducted experiments using four human cadaveric lumbar spine specimens and evaluate our algorithm in a cross validation experiment. Our method reaches a sensitivity of $91.50 \pm 6.58\%$ and a specificity of $91.10 \pm 2.27\%$. The proposed system shows great potentials for the development of alternative assessment methods for implant loosening based on vibroacoustic sensing.

1 Introduction

Spinal instrumentation with pedicle screws is an broadly established and increasingly used intervention in the surgical treatment of degenerative diseases, injuries, deformities or tumors of the spine. [1, 2, 3]. Hereby, the spinal segment is stabilized by driving screws into both pedicles of the respective vertebra and connect them with rods on either side that absorb most of the biomechanical forces. One of the most common postoperative complications of this surgical procedure is screw loosening, which is often associated with persistent pain and therefore eventually requires revision surgery. Pedicle screw loosening usually manifests itself in a fan-shaped cavity around the screw shaft and results in screw toggling, therefore allowing movement between the instrumented segments [4, 5, 6] The risk of pedicle screw loosening has been reported in literature as 1 – 3% per screw and 12.3% per patient [7]. In osteoporotic bone there is an even higher risk of pedicle screw loosening in the range of 50 – 60% [8, 9] which imposes a highly relevant clinical problem in an ageing population [10].

When patients report implant-related pain or instabilities, the standard way to asses potential implant loosening is to employ different medical imaging modalities, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), or planar radiographs. In a prospective clinical study, Spirig et al. found the sensitivity and specificity in detection of screw loosening to be 43.9% and 92.1% for MRI, 64.8% and 96.7%

for CT, and 54.2% and 83.5% for standard radiographs, respectively [11]. In clinical practice, CT remains the gold standard for the assessment of pedicle screw loosening but fails to detect a substantial amount of loose implants and exposes the patient to radiation, even though it is recommended to use low-dose CT protocols if possible [12]. Therefore, there is a clinical need to develop alternative non-invasive and radiation-free methods for the detection of loose pedicle screws and better understand their clinical correlation.

After screw loosening was diagnosed by imaging, a loose screw is confirmed intraoperatively by measuring a low torque when removing the screw [11, 13] which destroys the bone-implant interface in tight screws and further weakens the spine segment, if the screw was incorrectly identified as loose. Therefore, a reliable and non-invasive method for the intra-operative assessment of pedicle screw loosening for the intraoperative use in revision surgery would be highly desirable.

Acoustic sensing is a non-invasive, radiation-free and easy-to-integrate modality which has been shown to have great potential for various medical applications such as intraoperative tissue classification [14, 15], surgical error prevention [16], or patient monitoring [17]. Acoustic emission analysis has furthermore been employed in the condition assessment and early diagnosis of orthopedic implants, but has mainly been applied to artificial hip and knee joints so far [18]. Schwarzkopf et al. recorded the acoustic emissions of different types of knee implants using a handheld measurement system. The analysis of the data revealed correlations to the implant status and time from implantation [19]. Rodgers et al. proposed a system for monitoring the acoustic emissions of THA implants and characterized the squeaking of hard-on-hard bearing surface combinations [20]. Fitzpatrick et al. developed a monitoring system based on acoustic emission sensing to measure the wear of total hip replacement implants and compared the frequency characteristics of in-vivo and in-vitro recordings [21]. For the assessment of implant stability, a proof-of-concept-study was published by Ewald et al., who developed a prototype and simulator-based experimental setup for the detection of total hip replacement implant loosening using an acoustic sensor system [22]. Arami et al. developed a vibroacoustic system for ex vivo detection of loosening of total knee replacement implants. They applied harmonic vibration to the tibia and measured the resulting vibrations on the implant surface using an accelerometer [23].

The systems described above employ frequency analysis or classical signal processing methods to define thresholds or describe the characteristics of frequency components. However, as deep learning-based methods have recently replaced and outperformed classical approaches for solving audio specific tasks such as speech recognition [24] and environmental sound processing [25], these techniques have also successfully been applied to medical applications [17, 16, 26, 27].

In this work, we propose a novel method to assess the hold of pedicle screws based on vibroacoustic sensing. In the first step, we developed an experimental approach to simulate pedicle screw loosening in human cadaveric specimens. We instrumented four human cadaveric lumbar spine specimens and validated the screw loosening by analyzing the relative movement between implant and instrumented vertebra in fixed and loose configurations using a biomedical testing machine and an optical tracking system. For the detection of screw loosening, we excite the anatomy by using a vibration device to send a sine sweep into the bone and measure the propagated vibrations directly at the screw head. Subsequently, we developed an automated algorithm based on log-mel spectrograms and a SE-ResNet-18 to detect screw loosening based on the characteristics of the captured signal and thoroughly evaluate the performance of the proposed algorithm in a leave-one-specimen-out cross validation experiment. The proposed proof-of-concept system can be directly used for the intraoperative assessment of pedicle screw loosening during revision surgery.

2 Material and methods

2.1 Experimental approach for the simulation of Pedicle Screw Loosening

Pedicle screw loosening is usually simulated in biomechanical experiments by applying dynamic loading over thousands of cycles in experimental setups [4, 28, 29]. As we are not interested in measuring the biomechanical forces but rather simulate a loose implant in terms of relative movement between target anatomy and implant (screw toggling), we developed a different approach capable of simulating the mechanics of pedicle screw loosening in a faster way. To this end, CT scans of the cadaver specimens were acquired and the screw entry points and target angles were planned by an experienced spine surgeon according to the standard clinical routine in a 3D surgical planning software (CASPA, University Hospital Balgrist, Switzerland).

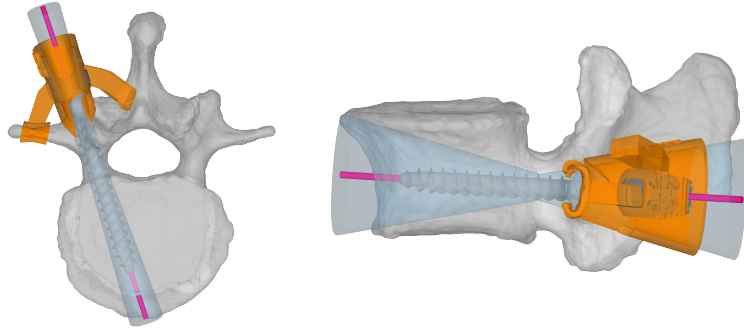


Figure 1: Instrumentation of the cadaveric spine specimens and simulation of the pedicle screw loosening using the 3D printed drill guide. On the left side, the sagittal view with an angle deviation of 25° is illustrated; the right side shows the transversal view with with an angle deviation of 5° . The custom 3D printed drill guide is colored in orange, the planned screw trajectory in pink, and the fan-shaped cavity in light blue.

A two-step data acquisition procedure was performed were first stable screws were inserted and measured (control group) and, in a second step, screw loosening was simulated (intervention group). In the first step, we inserted Medacta pedicle screws (Medacta, Castel San Pietro, Switzerland) along the planned screw trajectory using a classical approach without predrilling. In the second step, the screws were intentionally loosened. Therefore, we used the planned screw trajectories to design custom 3D printed drill guides, as illustrated in figure 1, to drill a fan-shaped hole with an angle deviation of 25° in the sagittal and 5° in the transversal plane into the respective vertebrae. The CTs were manually segmented and the drill guides leveraged the concept of patient specific instruments (PSI) [30] where the undersurface of the guide is shaped as a negative of the target anatomy surface, therefore only fitting in one unique position on the vertebrae. All custom drill guides were manufactured using a highly accurate laser sinter 3D printer (EOS Formiga P396, EOS Systems, Krailling, Germany).

For all four lumbar spine specimens, the vertebrae L2 and L4 were instrumented and only the screws in vertebra L2 were intentionally loosened in a second step. This approach allows to maintain a fixation on one implant side (L4) to measure the relative movement between implant and target vertebra (L2) during movement using an optical tracking system. All surgical steps were performed by an experienced spine surgeon and a bilateral posterior approach through the Wiltse interval was chosen to preserve the skin directly over the spinous processes.

2.2 Validation of Pedicle Screw Loosening

To validate the simulated screw loosening, we mounted each specimen in a biomechanical testing machine ZwickRoell Z010 (ZwickRoell GmbH & Co. KG, Ulm, Germany) which allows a defined and reproducible flexion-extension movement of the anatomy. To analyze the movement of the implants and the vertebrae, we attached passive optical tracking markers to the two rods, as well as to vertebrae L2 and L4. A high-fidelity optical tracking system, Atracsys fusionTrack 500 (Atracsys LLC, Puidoux, Switzerland), was used to record the trajectories of all tracked objects during movement. The experimental setup is shown in figure 2.

The protocol of the biomechanical testing machine was programmed with a maximal torque of 7.5 Nm in both directions, which determines the endpoint of the flexion-extension movement, and an angular speed of 5° s^{-1} . These values are standard settings for the biomechanical testing of the human lumbar spine according to empirical findings [31]. For the experimental validation of screw loosening, we run 30 - 50 cycles of flexion-extension movement and record the tracking data of the rigidly attached markers.

For all implants and specimens, screw loosening was confirmed by an experienced spine surgeon who conducted all experiments. As an additional experimental quantification metric for screw loosening, we introduce a ratio defined as the relation between fixed and loose configurations of the relative movement

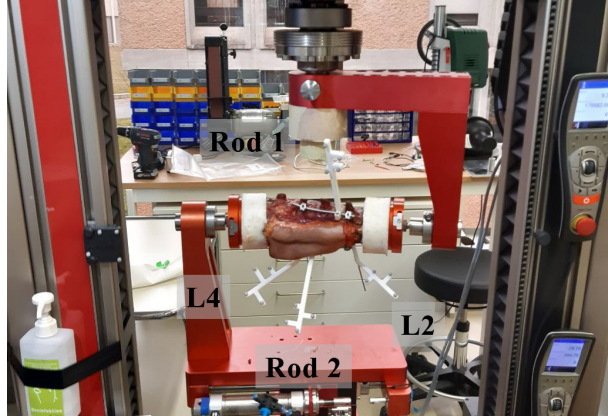


Figure 2: One of four cadaveric human lumbar spine specimens fixed in the biomedical testing machine. Individual passive infrared tracking markers are attached to vertebrae L2 and L4, as well as to the two rods of the implant.

between target vertebra and implant. In biomechanical screw toggling experiments, the level of screw loosening is usually measured as the relative displacement between implant and bone through optical tracking data as described in the work of Liebsch et al. [4]. However, these setups usually only include a single vertebra and screw mounted in a biomechanical testing machine which is only a simplified version of our setup. Furthermore, the proposed ratio compensates for subject-specific variations of relative displacement due to inter-subject bone quality differences.

First, we compute the relative movement Δx of each respective rod and the vertebra L2 for all configurations, where x is the reference point for each individual tracking target:

$$\Delta x_i = \|x_{L2} - x_{Rod_i}\| \quad (1)$$

The centered mean absolute relative movement is computed as a scalar measure of the amount of relative movement between the implant and the vertebra L2, where n is the number of synchronized measurements:

$$\bar{x}_i = \frac{1}{n} \sum_{i=1}^n |\Delta x_i - \frac{1}{n} \sum_{i=1}^n \Delta x_i| \quad (2)$$

Finally, we define the ratio of the relative movement between loose and fixed configuration as the loosening criterion and consider the screw as loose if the computed ratio exceeds a threshold of 2, which corresponds to a doubled relative movement of implant in regard to the target vertebra from fixed to loose configuration.

$$R_{lf,i} = \frac{\bar{x}_{i,loose}}{\bar{x}_{i,fixed}} > 2 \quad (3)$$

2.3 Vibro-Acoustic Sensing for Screw Loosening Detection

2.3.1 Experimental Setup

For the detection of pedicle screw loosening, we apply active vibration excitation to the target anatomy. We place a vibration device (shaker type 4810, Brüel and Kjær, Teknikerbyen 28, DK-2830 Virum, Denmark) on the skin centered on top of the spinous process of the target vertebra L2 in an upright position and excite the tissue with a sine sweep in a frequency range of 10 to 500 Hz and with a duration of 2.5 s. To standardize the applied pressure and to make the system easily usable, we use the weight of the shaker (1080 g) to define the contact pressure on the skin of the specimen and hold it manually in place. We use a digital oscilloscope, Digilent Analog Discovery 2 (Digilent, 1300 NE Henley Ct. Suite 3, Pullman, WA 99163, USA), and its MATLAB (MathWorks, 1 Apple Hill Drive, Natick, MA, USA) interface to generate the excitation signal.

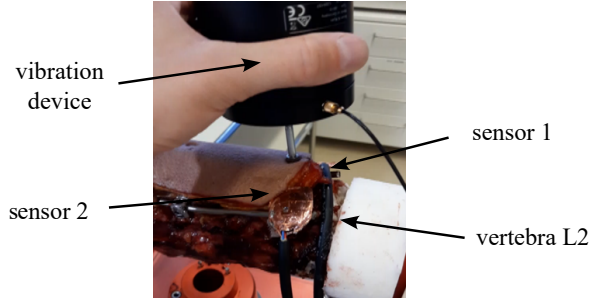


Figure 3: The vibration device is placed on top of the spinous process of the vertebra L2. The structure-borne vibrations are propagated through the bone into the screw shaft and measured with custom piezo-based contact microphones directly at the screw head.

We amplify (type 2706, Brüel and Kjær) the generated analog signal to drive the shaker. Figure 3 illustrates the experimental setup for the vibration experiments.

The vibration is propagated through the bony tissue of the vertebra to the screw shaft and recorded with custom piezo contact microphones which are directly glued to the screw head. The piezo contact microphones use a custom preamplification and impedance buffering stage to preserve low frequency content as described in [16]. We recorded a total number of 50 sine sweeps per screw and lifted and replaced the shaker device between the individual sweeps to have variation in the captured training data. We first measure every specimen in fixed configuration as illustrated in section 2.2, afterwards we intentionally loosen the screws of vertebra L2 using the protocol described in section 2.1 and repeat the whole procedure which results in 200 samples recorded per specimen. Using four human cadaveric lumbar spine specimens, we recorded a balanced dataset with a total number of 800 individual samples. All signals were captured in lossless wave file format, using a sample rate of 44.1 kHz and a bit depth of 24.

2.3.2 Pedicle Screw Loosening Detection Algorithm

State-of-the-art systems in audio classification use a combination of log-mel spectrogram representation for the audio signal and a feature extraction backbone based on convolutional neuronal networks [33, 34]. Therefore, we generate log-mel spectrograms with dimensions 256x218 from all individual samples in the dataset. The log-mel spectrograms serve as input for a modified 18-layer ResNet [35] and are computed using the python library librosa 0.7.2 [36]. Log-mel spectrograms are two-dimensional matrices with time windows as columns, frequency mel-bins as rows, and amplitude as scalar matrix values. The first step to compute the log-mel spectrogram of an audio sample of length N is to compute Short-Time Fourier Transformation (STFT):

$$X(m, k) = \sum_{n=0}^{N-1} x(n + mH)w(n)exp\left(\frac{-2\pi ikn}{N}\right) \quad (4)$$

We use the Hann window function as $w(n)$ to compensate for spectral leakage [37] and apply a hop length of $H = 256$. We map the resulting STFT X which is structured as the k^{th} Fourier coefficient (on the y-axis) for the m^{th} time frame (on the x-axis) from amplitude to decibel by computing:

$$X_{dB}(m, k) = 10 \log_{10}(X(m, k)^2) \quad (5)$$

Finally, the spectrogram is mapped to the mel scale by applying a total number of 256 triangular filters which are evenly distributed on the Mel scale defined by:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (6)$$

Compared to previous work, we extended the ResNet-18 backbone with Squeeze & Excitation (SE) [32] blocks which add a channel-wise attention mechanism to each residual block while introducing minimal

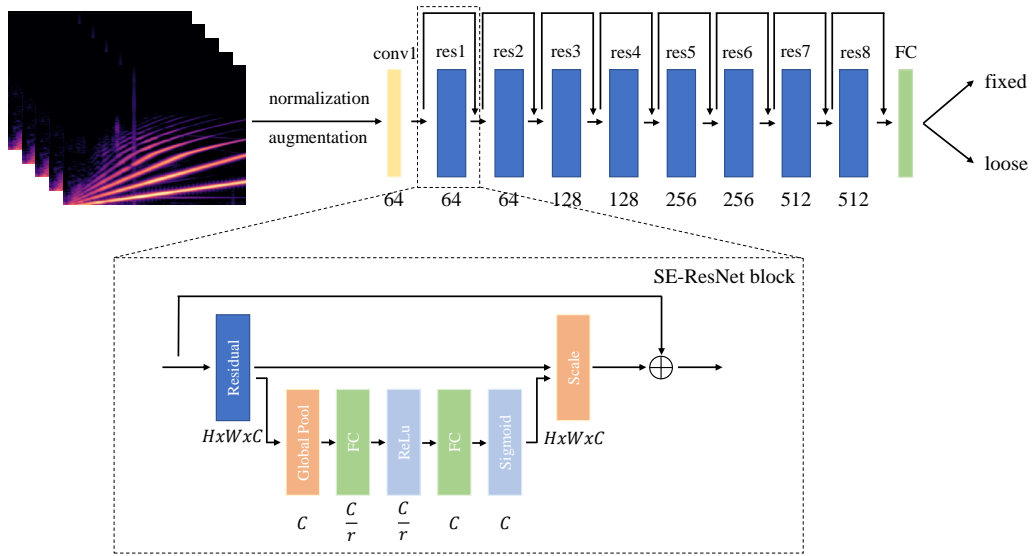


Figure 4: The overview of the proposed pedicle screw loosening detection pipeline. The spectrograms are fed to an 18 layer ResNet variant which implements Squeeze & Excitation [32] in each residual block. We define the detection as a binary classification problem, where the model is trained to differentiate between fixed and loose samples. In the SE-ResNet block schematic, the dimensions of each layer are illustrated where the variable r stands for the reduction ratio as described in [32]. The number of filters for each SE-ResNet block are given below, every layer employs a filter size of 3×3 . The spectrograms are colored for visualization purposes, however all spectrogram used in the implementation of this work are two-dimensional matrices with dimensions 256×218 .

ID	Screw1 fixed	Screw1 loose	Screw2 fixed	Screw2 loose
0	0.5998 mm	2.5323 mm	0.9927 mm	3.3198 mm
1	0.5122 mm	1.3140 mm	0.2181 mm	1.5746 mm
2	0.9234 mm	2.0103 mm	0.6141 mm	6.2079 mm
3	1.2149 mm	3.3013 mm	0.4763 mm	4.8975 mm

Table 1: The relative displacement of implant and target vertebra.

ID	$R_{lf,screw1}$	$R_{lf,screw2}$
0	6.9462	4.2048
1	4.7403	28.3358
2	2.6498	14.1524
3	3.1605	15.5463

Table 2: The values of $R_{lf,screw1}$ and $R_{lf,screw2}$ for all specimens and screws. The ratios have been computed according to the measurements and formulas described in sections 2.1 and 2.2.

computational overhead. This modification resulted in a substantial performance improvement in our experiments. In our implementation, we use a reduction ratio of $r = 8$ in all SE blocks. An overview of the proposed detection pipeline is illustrated in figure 4.

The network is trained for 10 epochs using the Adam optimizer, a learning rate of $L_R = 1e - 5$, and a binary crossentropy loss. We normalize all log-mel spectrograms $X_{norm,mel} = (X_{mel} - \mu)/\sigma$, where μ is overall mean and σ is the standard deviation computed over the entire dataset. Furthermore, we augment the dataset by applying pitch shifting in the range of $[-3, 3]$ semitones and time stretching with the factors $[0.9, 1.1]$ to the raw waveforms directly. All experiments were implemented in Tensorflow/Keras 2.6 and executed on a NVidia RTX 2080 SUPER GPU.

3 Results

3.1 Validation of Screw Loosening

Table 1 shows the relative displacement of implant and vertebra L2 measured using the optical tracking system for every implant and specimen in both fixed and loose configurations. It can be observed that every specimen and implant shows an increased displacement after intentional loosening. To compensate for inter-subject variations in relative displacement due to differences in bone quality, anatomy and mounting in the biomechanical testing machine, we computed the loosening ratio for each individual screw as described in section 2.2 as the main metric for the assessment of screw loosening. Table 2 contains the computed ratios for all screws and specimens tested in our experiment.

3.2 Screw Loosening Detection Results

The spectrograms are analyzed using the detection pipeline proposed in section 2.3.2. To thoroughly evaluate the model performance, we perform a four-fold cross validation experiment and split the data on a specimen level. To this end, we train an individual model from scratch on the data collected from three specimens and test on the remaining specimen. All results are reported in the format *mean \pm standard deviation*.

For the detection of pedicle screw loosening, our model reaches a sensitivity of $91.50 \pm 6.58\%$ and a specificity of $91.10 \pm 2.27\%$. These values correspond to a mean accuracy of the detection algorithm of $91.29 \pm 4.28\%$. To give further insights, we report the results on the individual folds in table 3, where the specimen ID corresponds to the specimen used for the test set, the model is trained on the remaining three specimens.

We furthermore performed an ablation study to show the benefit of modifying the ResNet-18 backbone with Squeeze & Excitation modules for the given problem. Without Squeeze & Excitation modules, the

Specimen ID	Sensitivity	Specificity
0	94.00%	94.00%
1	81.00%	87.76%
2	92.00%	90.62%
3	99.00%	92.00%

Table 3: Sensitivity and specificity reported for each individual fold in the four-fold cross validation experiment conducted for the evaluation of the proposed pedicle screw detection algorithm.

model reached a sensitivity of $87.75 \pm 9.91\%$ and a specificity of $90.04 \pm 7.19\%$ which corresponds to a mean accuracy of $88.89 \pm 4.08\%$.

4 Discussion

In this work, we propose a detection method for pedicle screw loosening based on vibroacoustic sensing which could be an important step towards a novel radiation-free and non-invasive assessment method to improve the diagnostics in clinical practice and patient safety in revision surgery. We thoroughly evaluate our algorithm using k-fold cross validation and split the dataset on the specimen level. To the best knowledge of the authors, we propose the first alternative to medical imaging based assessment methods. Our experimental design may also allow clinical translation to a percutaneous application with reproduction of the typical pain in the context of symptomatic screw loosening. This would help physicians as the clinical correlation of radiological findings of screw loosening with the complained symptoms is not always evident. To address the aforementioned problem, the proposed learning-based pedicle screw loosening detection algorithm shows promising performance indicating great potential for the development of systems for the automated screw loosening detection based on vibroacoustics. As the target vertebra is excited with sine sweep vibration, the resulting measurements at the screw head are greatly influenced by the anchorage of the screw in the surrounding bone tissue. A fan-shaped cavity around the screw shaft therefore changes the transmitted vibration characteristics which serves as the basic structure-borne sound propagation mechanism that motivates our work.

3D-printed surgical guides were introduced as an approach for a more time efficient simulation of pedicle screw loosening which was confirmed to be sufficiently realistic through the analysis of optical tracking data. Extended simulation with toggling experiments would probably result in a more realistic screw loosening model as the loosening funnel was uniformly designed for all specimens. However, the focus of the present work is the development of a vibroacoustic-based method for screw loosening detection, the implementation of a highly realistic loosening simulation is not in the scope of this work and should be investigated in future research. Furthermore, as we fully loosened the screw in our experiments, the influence on the detection performance of the proposed algorithm with different levels of screw loosening has to be investigated in future work. In addition, the influences of patient body mass index (BMI) and bone quality on the proposed system should be taken into consideration.

After conducting the loosening simulation process as described in section 2.1, an experienced spine surgeon confirmed the loosening of the respective pedicle screws visually and haptically. To additionally quantify the screw loosening, we introduced a loosening ratio which is computed using optical tracking data. Hereby, the relative displacement of implant and target vertebra shown in 1, as well as the loosening ratio shown in table 2 show certain variations. The reasons for these variations are subject-specific bone quality and anatomical differences. Furthermore, it is practically unfeasible to install the specimen perfectly centered in the biomechanical testing machine which results in a slightly asymmetric movement. However, the loosening validation experiments show a more than doubled relative movement between implant and respective vertebra which can be considered sufficient for simulating pedicle screw loosening in our experiments.

A limitation of the presented study is the small sample size of four human cadaveric specimens. However, by performing a four-fold cross validation experiment and showing the consistency of the results over all four individual test folds, we consider our experiment as a strong proof-of-concept for the usage of vibroacoustic sensing in orthopedics. The variations in the per-fold model performance can be accounted to anatomical variations and bone qualities. Additional *in-vitro* and *in-vivo* studies have to be performed to test the

reliability of the system and increase the training data for better generalization and detection performance.

With a shaker device that applies the vibration on the patient’s skin over the spinous process to the bone and using its weight as contact force, we propose an easy-to-integrate measurement method which requires only little additional human effort. In the present work, we chose a bilateral posterior approach for the surgical access, however, also with a central surgical access, the method is suitable for intraoperative detection of loose pedicle screws, as the vibration device cannot only be placed on top of the spinous process on the skin, but also directly on the bone. In future, we envision the presented approach not only to be valuable as an intraoperative confirmation of pedicle screw loosening, but also as a clinical tool for the preoperative diagnosis of screw loosening and, eventually, as the foundation to design smart pedicle screws to monitor or even predict pedicle screw loosening in a reliable and non-invasive way. Nevertheless, additional research and development is required to design custom sensorized implants, transmit the signals to the outside of the human body and solve the energy supply.

5 Conclusions

We propose a non-destructive, radiation-free and easy-to-integrate approach to detect pedicle screw loosening intraoperatively using active vibroacoustic sensing. The resulting system could be used for the intraoperative confirmation of loose pedicle screws as an alternative for the measurement of the extractional torque. Furthermore, we believe that the proposed work could be a strong proof-of-concept for the development of smart implants for spinal fusion surgery.

Acknowledgments

This work is part of the SURGENT project and was funded by University Medicine Zurich/Hochschulmedizin Zürich.

References

- [1] R. J. Mobbs, K. Phan, G. Malham, K. Seex, P. J. Rao, Lumbar interbody fusion: techniques, indications and comparison of interbody fusion options including plif, tlif, mi-tlif, olif/atp, llif and alif, *Journal of Spine Surgery* 1 (2015) 2–18.
- [2] S. S. Rajaei, H. W. Bae, L. E. A. Kanim, R. B. Delamarter, Spinal fusion in the united states: analysis of trends from 1998 to 2008, *Spine* 37 (2012) 67–76.
- [3] K. Kobayashi, K. Sato, F. Kato, T. Kanemura, H. Yoshihara, Y. Sakai, R. Shinjo, T. Ohara, H. Yagi, Y. Matsubara, K. Ando, H. Nakashima, S. Imagama, Trends in the numbers of spine surgeries and spine surgeons over the past 15 years, *Nagoya Journal of Medical Science* 84 (2022) 155—162.
- [4] C. Liebsch, J. Zimmermann, N. Graf, C. Schilling, H.-J. Wilke, A. Kienle, In vitro validation of a novel mechanical model for testing the anchorage capacity of pedicle screws using physiological load application, *Journal of the Mechanical Behavior of Biomedical Materials* 77 (2018) 578–585.
- [5] M. Law, A. F. Tencer, P. A. Anderson, Caudo-cephalad loading of pedicle screws: mechanisms of loosening and methods of augmentation, *Spine* 18 (1993) 2438–2443.
- [6] D. A. Baluch, A. A. Patel, B. Lullo, R. M. Havey, L. I. Voronov, N.-L. Nguyen, G. Carandang, A. J. Ghanayem, A. G. Patwardhan, Effect of physiological loads on cortical and traditional pedicle screw fixation, *Spine* 39 (2014) 1297–1302.
- [7] J. Bredow, C. K. Boese, C. M. L. Werner, J. Siewe, L. Löhner, K. Zarghooni, P. Eysel, M. J. Scheyerer, Predictive validity of preoperative ct scans and the risk of pedicle screw loosening in spinal surgery, *Archives of Orthopaedic and Trauma Surgery* 136 (2016) 1063–1067.

- [8] A. El Saman, S. Meier, A. Sander, A. Kelm, I. Marzi, H. Laurer, Reduced loosening rate and loss of correction following posterior stabilization with or without pmma augmentation of pedicle screws in vertebral fractures in the elderly, *European Journal of Trauma and Emergency Surgery* 39 (2013) 455–460.
- [9] J. H. Kim, D. K. Ahn, W. S. Shin, M. J. Kim, H. Y. Lee, Y. R. Go, Clinical effects and complications of pedicle screw augmentation with bone cement: Comparison of fenestrated screw augmentation and vertebroplasty augmentation, *Clinics in Orthopedic Surgery* 12 (2020) 194–199.
- [10] S. Hoppe, M. J. B. Keel, Pedicle screw augmentation in osteoporotic spine: indications, limitations and technical aspects, *European Journal of Trauma and Emergency Surgery* 43 (2017) 3–8.
- [11] J. M. Spirig, R. Sutter, T. Götschi, N. A. Farshad-Amacker, M. Farshad, Value of standard radiographs, computed tomography, and magnetic resonance imaging of the lumbar spine in detection of intraoperatively confirmed pedicle screw loosening—a prospective clinical trial, *The Spine Journal* 19 (2019) 461–468.
- [12] K. Abul-Kasim, A. Ohlin, Evaluation of implant loosening following segmental pedicle screw fixation in adolescent idiopathic scoliosis: a 2 year follow-up with low-dose ct, *Scoliosis* 9 (2014).
- [13] X. Wu, J. Shi, J. Wu, Y. Cheng, K. Peng, J. Chen, H. Jiang, Pedicle screw loosening: the value of radiological imagings and the identification of risk factors assessed by extraction torque during screw removal surgery, *Journal of Orthopaedic Surgery and Research* 14 (2019).
- [14] D. Ostler, M. Seibold, J. Fuchtmann, N. Samm, H. Feussner, D. Wilhelm, N. Navab, Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions, *International Journal of Computer Assisted Radiology and Surgery* (2020).
- [15] A. Illanes, A. Boese, I. Maldonado, A. Pashazadeh, A. Schaufler, N. Navab, M. Friebe, Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing, *Scientific Reports* 8 (2018).
- [16] M. Seibold, S. Maurer, A. Hoch, P. Zingg, M. Farshad, N. Navab, P. Fürnstahl, Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery, *Scientific Reports* 11 (2021).
- [17] H. E. Romero, N. Ma, G. J. Brown, A. V. Beeston, M. Hasan, Deep learning features for robust detection of acoustic events in sleep-disordered breathing, in: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 810–814.
- [18] L. Khokhlova, D.-S. Komaris, S. Tedesco, B. O’Flynn, Assessment of hip and knee joints and implants using acoustic emission monitoring: A scoping review, *IEEE Sensors* 21 (2021) 14379–14388.
- [19] R. Schwarzkopf, F. J. Kummer, W. L. Jaffe, Acoustic emission studies of posterior stabilized and cruciate retaining knee arthroplasties, *Journal of Knee Surgery* 24 (2011) 185–189.
- [20] G. W. Rodgers, J. L. Young, A. V. Fields, R. Z. Shearer, T. B. F. Woodfield, G. J. Hooper, J. G. Chase, Acoustic emission monitoring of total hip arthroplasty implants, *IFAC Proceedings Volumes* 47 (3) (2014) 4796–4800, 19th IFAC World Congress.
- [21] A. J. Fitzpatrick, G. W. Rodgers, G. J. Hooper, T. B. Woodfield, Development and validation of an acoustic emission device to measure wear in total hip replacements in-vitro and in-vivo, *Biomedical Signal Processing and Control* 33 (2017) 281–288.
- [22] H. Ewald, U. Timm, C. Ruther, W. Mittelmeier, R. Bader, D. Kluess, Acoustic sensor system for loosening detection of hip implants, in: *2011 Fifth International Conference on Sensing Technology*, 2011, pp. 494–497.
- [23] A. Arami, J.-R. Delaloye, H. Rouhani, B. M. Jolles, K. Aminian, Knee implant loosening detection: A vibration analysis investigation, *Annals of Biomedical Engineering* 46 (2018) 97–107.

- [24] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, Speech recognition using deep neural networks: A systematic review, *IEEE Access* 7 (2019) 19143–19165.
- [25] K. Miyazaki, T. Toda, T. Hayashi, K. Takeda, Environmental sound processing and its applications, *Transactions on Electrical and Electronic Engineering* 14 (2019) 340–351.
- [26] X. Xu, E. Nemati, K. Vatanparvar, V. Nathan, T. Ahmed, M. M. Rahman, D. McCaffrey, J. Kuang, J. A. Gao, Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5 (1) (2021).
- [27] M. Seibold, A. Hoch, D. Suter, M. Farshad, P. O. Zingg, N. Navab, P. Furnstahl, Acoustic-based spatio-temporal learning for press-fit evaluation of femoral stem implants, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 447–456.
- [28] T. J. Choma, W. F. Frevert, W. L. Carson, N. P. Waters, F. M. Pfeiffer, Biomechanical analysis of pedicle screws in osteoporotic bone with bioactive cement augmentation using simulated in vivo multicomponent loading, *Spine* 36 (2011) 454–462.
- [29] R. A. Kueny, J. P. Kolb, W. Lehmann, K. Puschel, M. M. Morlock, G. Huber, Influence of the screw augmentation technique and a diameter increase on pedicle screw fixation in the osteoporotic spine: pullout versus fatigue testing, *European Spine Journal* 23 (2014) 2196–2202.
- [30] M. A. Hafez, K. Moholkar, Patient-specific instruments: advantages and pitfalls, *SICOT-J* 6 (2017).
- [31] H. J. Wilke, K. Wenger, C. L. Testing criteria for spinal implants: recommendations for the standardization of in vitro stability testing of spinal implants, *European Spine Journal* 7 (1998) 148–154.
- [32] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [33] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-y. Chang, T. Sainath, Deep learning for audio signal processing, *IEEE Journal on Selected Topics in Signal Processing* 14 (2019) 206–219.
- [34] D. Ćirić, Z. Perić, J. Nikolić, N. Vućić, Audio signal mapping into spectrogram-based images for deep learning applications, in: *2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2021, pp. 1–6.
- [35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [36] B. McFee, C. Raffel, D. Liang, D. PW Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python., in: *14th python in science conference*, 2015, pp. 18–25.
- [37] D. A. Lyon, The discrete fourier transform, part 4: Spectral leakage, *Journal of Object Technology* 8 (7) (2009) 23–34.

3.3 Spectrogram-based Spatio-temporal Learning

In spatio-temporal learning, time information in the form of sequential data is incorporated in the modeling of the problem in addition to a frame-based processing of data points. Even though a spectrogram captures a signal over a certain predetermined time frame, spatio-temporal learning enables the analysis of longer data sequences. While each data point, e.g. a spectrogram, is spatially processed during feature extraction, subsequent sequential layers can be employed to incorporate the sequential nature of the temporal succession of these extracted features from each individual spectrogram. The temporal modeling can be realized using different techniques, e.g. through Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) layers, or more recently self-attention mechanisms.

3.3.1 Contribution (Interventional): Acoustic-Based Spatio-Temporal Learning for Press-Fit Evaluation of Femoral Stem Implants (MICCAI 2021)

Summary: In Total Hip Arthroplasty (THA), the femoral head is resected and increasing sizes of broaches are driven into the femur with hammer blows to prepare the femoral canal for the final implant and ensure a good press-fit between bone and implant. One of the most common complications during this step are periprosthetic fractures which cause severe pain and increased trauma for patients undergoing THA, resulting in extended rehabilitation. Furthermore, periprosthetic fractures can, in the worst case, remain undetected which requires additional revision surgery. Because conventional surgical navigation systems are not able to assess the level of press-fit and implant seating and surgeons reported to use the changing hammer blow sounds during insertion for implicit guidance, acoustic signals have great potential as an alternative sensing modality to assess the implant seating based on acoustic characteristics of the hammer blow events. In this work, we propose to sensorize the inserter tool using a high sensitive contact microphone to capture structure-borne vibration characteristics from the broach-inserter structure during insertion. We furthermore introduce a spatio-temporal model which is trained to classify sequences of five hammer blow events into increasing and reached target press fit based on a preoperative plan of the procedure. The proposed system is evaluated in a human cadaveric experimental setup using a five-fold cross validation scheme. Our results show great potential for the design of acoustic sensing-based system for error prevention and intraoperative decision support systems in hip surgery.

Contributions: The author of this thesis was responsible for formulating the problem and approach with medical consultations from Armando Hoch, Daniel Suter, Mazda Farshad, and Patrick Zingg as well as for capturing the data in the experimental setup, designing the data processing, detection, and evaluation pipeline and writing the manuscript. All human cadaveric experiments were conceptualized and performed by the author and Armando Hoch. Nassir Navab and Philipp Frnstahl contributed in the form of discussions and feedback throughout the whole project and for proofreading of the manuscript.

Copyright Statement: Reprint of the final author's accepted manuscript by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2021, Lecture Notes in Computer Science, vol 12904, Matthias Seibold, Armando Hoch, Daniel Suter, Mazda Farshad, Patrick O. Zingg, Nassir Navab & Philipp Fürnstahl, "Acoustic-Based Spatio-Temporal Learning for Press-Fit Evaluation of Femoral Stem Implants" (2021).

The published version of this work can be accessed under: https://doi.org/10.1007/978-3-030-87202-1_43

Acoustic-based Spatio-temporal Learning for Press-fit Evaluation of Femoral Stem Implants

Matthias Seibold^{1,2}, Armando Hoch³, Daniel Suter³, Mazda Farshad³, Patrick Zingg³, Nassir Navab¹, and Philipp Furnstahl^{2,3}

¹ Computer Aided Medical Procedures (CAMP), Technical University of Munich, Munich, DE-85748, Germany

² Research in Orthopedic Computer Science (ROCS), University Hospital Balgrist, University of Zurich, Balgrist CAMPUS, Zurich, CH-8008, Switzerland

³ Balgrist University Hospital, Zurich, CH-8008, Switzerland

Abstract. In this work, we propose a method utilizing tool-integrated vibroacoustic measurements and a spatio-temporal learning-based framework for the detection of the insertion endpoint during femoral stem implantation in cementless Total Hip Arthroplasty (THA). In current practice, the optimal insertion endpoint is intraoperatively identified based on surgical experience and dependent on a subjective decision. Leveraging spectrogram features and time-variant sequences of acoustic hammer blow events, our proposed solution can give real-time feedback to the surgeon during the insertion procedure and prevent adverse events in clinical practice. To validate our method on real data, we built a realistic experimental human cadaveric setup and acquired acoustic signals of hammer blows during broaching the femoral stem cavity with a novel inserter tool which was enhanced by contact microphones. The optimal insertion endpoint was determined by a standardized preoperative plan following clinical guidelines and executed by a board-certified surgeon. We train and evaluate a Long-Term Recurrent Convolutional Neural Network (LRCN) on sequences of spectrograms to detect a reached target press fit corresponding to a seated implant. The proposed method achieves an overall per-class recall of $93.82 \pm 5.11\%$ for detecting an ongoing insertion and $70.88 \pm 11.83\%$ for identifying a reached target press fit for five independent test specimens. The obtained results open the path for the development of automated systems for intra-operative decision support, error prevention and robotic applications in hip surgery.

Keywords: Spatio-temporal Learning · Acoustic Sensing · Total Hip Arthroplasty · Femoral Stem Insertion.

1 Introduction

For the preparation of the femur for implant insertion in cementless THA, the femoral head is resected and broaches of increasing size are driven into the femoral stem cavity with hammer blows, before the final femoral stem implant is inserted. A frequent intraoperative complication during this procedure

is periprosthetic femoral fracture which has been reported to occur with rates of 3.5% to 5.4% [2,3,19]. Hereby, the majority of periprosthetic fractures (46.5%) happen during the preparation of the femur for stem insertion [1] and are mainly caused by excessive broaching beyond the optimal insertion endpoint. Figure 1 shows a radiograph of a periprosthetic fracture. A relevant part of these fractures, so-called occult periprosthetic fractures, cannot even be discovered intraoperatively [23]. The standard clinical procedure to assess the seating of the femoral stem implant intraoperatively and determine the optimal insertion endpoint is based on preoperative planning, surgical experience, simple distance measurements, and radiologic verification of the implant seating in radiographs [10]. Therefore, a system which is able to assess the seating of the implant in the femoral cavity, inform the surgeon when the target press-fit is reached, and reduce the risk of intraoperative periprosthetic fracture would be highly desirable.

An ideal system would detect the optimal insertion endpoint during broaching and inform the surgeon to stop the insertion procedure. Conventional navigation systems can provide support in finding the correct implant position [18], but they cannot guide the surgeon in finding the insertion endpoint. As the underlying problem is not geometric, we chose to focus on a different data modality. Acoustic signals have been shown to be a rich source of information in medical scenarios, for example for the applications in bone drilling [20], arthroscopy [21], or needle injection [9]. Also for the assessment of the insertion endpoint in femoral stem insertion, acoustic signals which are generated by the impact of the hammer onto the broach inserter have been analyzed in prior work. For femoral stem insertion, hammer blow sounds have been identified to be correlated to complications by Morohashi et al. [14]. A cadaveric study was performed by Oberst et al. [15] and a nonlinear time-series analysis of the impulse response function showed relations to the process of femoral broaching. However, no clear insertion endpoint or stopping criterion was identified in this work.

Goossens et al. identified a number of hand-crafted features which correlate with the implant seating in *in-vitro* [5] and *in-vivo* [6] experiments using air-borne microphones and defined a stopping criterion based on the convergence of these features. However, air-borne sensors have the disadvantage of capturing environmental noise of the operating room in contrast to contact-based vibroacoustic measurements. Furthermore, they did not implement a system for the automated assessment of the implant seating. Additionally, learned features have been shown to have advantages over handcrafted features, such as better generalization and better performance with an increasing amount of training data [11].

A sensorized instrument for the monitoring of cementless femoral stem insertion has been proposed by Tijou et al. [22] employing a force sensor attached to the impacting face of the hammer to measure the impact force. This preliminary study using artificial bone models was later validated in a cadaveric experiment [4]. They showed that a time-domain and peak-based feature shows correlations to the displacement of the implant, measured using optical markers. Also in this

work, no automated system was developed and features were handcrafted and not learned.

In this work, we propose a method which enables smart instruments for orthopedic surgery through tool-integrated vibroacoustic measurements and developed a deep learning framework for the automated analysis of hammer blow sounds in THA. By not attaching the piezo element to the hammer, but to the inserter tool itself, we capture the full vibration response of the broach-inserter structure instead of only the impact force. We furthermore advance the state-of-the-art in vibroacoustic analysis of surgical procedures by introducing a spatio-temporal model for sequence-based press fit evaluation and show the benefits of including time-domain data in our framework. We thoroughly evaluate the performance of our model in cadavers using 5-fold cross validation. In the following sections we introduce our method and describe the experimental setup and data generation process.

2 Materials and Method

2.1 Data Pre-processing and Spatio-temporal Model for Press-fit Evaluation

To capture acoustic signals of the surgical procedure, we used a custom piezo-based contact microphone and attached it to the inserter tool to capture structure-borne sounds of the hammer blows and the resonance of the tool during the insertion process. The piezo-sensor is electromagnetically shielded and impedance-buffered to minimize noise and optimize the frequency transmission. We attach the sensor to the tool using electrical tape for firm fixation. The inserter with attached contact sensor is illustrated in Figure 1.

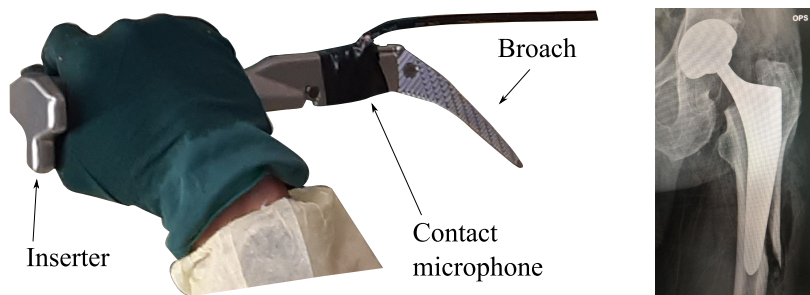


Fig. 1. Left: The inserter-broach structure with attached piezo-based contact microphone. Right: A radiograph of a periprosthetic fracture.

From the recorded soundclips, we extract spectrogram features which are lower-dimensional compared to raw audio and have been shown to yield promising performance for general audio signal processing [17], but also for automated

acoustic analysis in surgical applications [16,20]. Log-mel-spectrograms, a sparse and high resolution spectrogram variant, with dimensions 256x344x1 were computed from the individual sound clips of hammer blows using the python library *librosa 0.7.2* [13]. Log-mel spectrograms are two-dimensional matrices with time windows as columns, frequency mel-bins as rows, and amplitude as scalar matrix values. First, Short-Time Fourier Transformation (STFT) was applied to each sound clip of length N using:

$$X(m, k) = \sum_{n=0}^{N-1} x(n + mH)w(n)exp\left(\frac{-2\pi ikn}{N}\right) \quad (1)$$

In our implementation, the term $w(n)$ corresponds to the *Hann* window function to avoid *spectral leakage* [12]. The resulting STFT spectrogram X contains the k^{th} Fourier coefficient for the m^{th} time frame as matrix values. We used a hop length of $H = 16$ for all spectrograms. We computed the power spectrogram from X and mapped it to the logarithmic decibel scale using:

$$X_{pow}(m, k) = 10 \log_{10}(X(m, k)^2) \quad (2)$$

To finally convert this log STFT spectrogram to the mel scale, a total number of 256 triangular filters which are spaced evenly on the Mel scale (equation 3) were applied to X_{pow} . The mel scale can be computed from frequency by:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (3)$$

We normalized all spectrograms $X_{norm,mel} = (X_{mel} - \mu)/\sigma$, where (μ) is overall mean and (σ) is the overall standard deviation.

The network architecture consists of a ResNet-50 backbone [7], an architecture which has been shown to perform especially well for audio classification tasks [8,20], which extracts features from each input spectrogram in the sequence. The input of the proposed model is a sequence of five spectrograms of consecutive hammer blows. The output of the time-distributed, randomly initialized ResNet is a sequence of five feature vectors of size 1x2048 which is passed to an LSTM layer and subsequently to two consecutive fully connected layers. In between the two fully connected layers, we apply a dropout of 0.5 to reduce the model's tendency towards overfitting. The entire model has a total of 26,050,945 parameters and outputs the probabilities to classify the sequences into the classes $C := \{\text{insertion, press fit}\}$, where c_i denotes the respective class. An outline of the proposed detection pipeline is illustrated in Figure 2. We trained the model using a batch size of eight sequences and the Adam optimizer with an empirically determined optimal learning rate of $1 * 10^{-5}$, minimizing a binary crossentropy loss. Early stopping is used for additional regularization. Data augmentation was applied during training using time-stretching by a factor of [0.5, 0.7, 1.2, 1.5] and pitch-shifting by [-3, -1, +1, +3] semitones on the audio file level. All experiments were performed using Tensorflow/Keras 2.2 on a NVIDIA GeForce RTX 2080 SUPER GPU. The implementation is available under https://caspa.visualstudio.com/CARD%20public/_git/AudioFemoralStem.

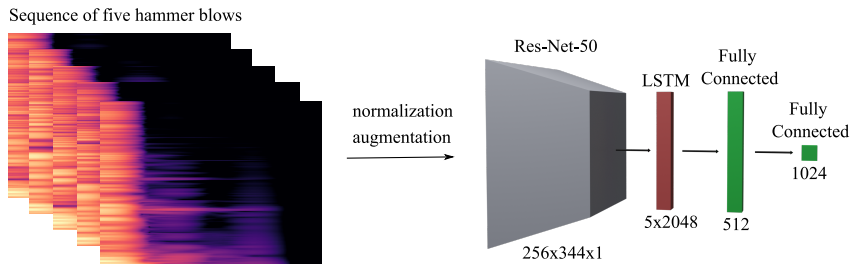


Fig. 2. The outline of the proposed spatio-temporal model. The input sizes of each layer of the pipeline is stated at the bottom. For visualization purposes, the spectrograms are displayed as color images, however the implementation uses features of size 256x344x1.

2.2 Experimental Setup and Data Generation

To build an experimental setup, which is as close as possible to the clinical scenario, we conducted experiments with five thawed fresh-frozen human cadaveric hip specimens including soft tissue. An ethical approval and informed consent from all study subjects was obtained. One orthopedic surgeon prepared the femur using a standard anterior approach. To follow the clinical procedure for femoral stem insertion, we broached the femoral stem cavity using the Medacta *QUADRA* system (Medacta International SA, Castel San Pietro, Switzerland), consisting of broaches of increasing size. The target broach size was planned using the clinical planning software *mediCAD* (mediCAD Hectec GmbH, Altdorf/Landshut, Germany) and the surgical plan was executed by a trained surgeon for each specimen. The outline of the preoperative plan and the cadaveric experiments is illustrated in the last column of table 1. The insertion endpoint was classified as the preoperatively planned broach size fully seated in the anatomy and confirmed by the surgeon.

Structure-borne acoustic signals were acquired from the contact sensor attached to the tool during the broaching process. We additionally captured video footage of the whole experiment to facilitate the labelling process. For the generation of the training data set, the recorded sequences of hammer blows were labelled into two classes, $\{c_0, c_1\}$. The class c_0 contains audio samples of hammer blows during insertion of the increasing sizes of broaches until the planned target size. The operating surgeon identified the target press fit as the preoperatively planned broach size fully seated in the anatomy. The class c_1 contains samples of hammer blows after reaching the target press fit.

We use the PreSonus Studio 68 (PreSonus Audio Electronics, Inc., Baton Rouge, LA, USA) audio interface and the Audio Stream Input/Output (ASIO) low-latency driver to capture loss-less audio with a sampling rate of 44.1 kHz and a bit depth of 24 bit. All samples were cut to a length of $N = 5500$ samples, which corresponds to a duration of 125 ms which has been empirically determined as the optimal length to capture the whole duration of a hammer blow sound.

During real surgery, these hammer blows could be detected by a simple threshold in the recorded structure-borne audio. For 5-fold cross validation, we split the data set on the specimen level and use the data from four specimens for training and from one specimen for testing, respectively. The final data set contains a total number of 1795 ($n_{c_0} = 1245$, $n_{c_1} = 550$) sequences of five hammer blows.

3 Results and Evaluation

In the following sections we present the evaluation of the proposed spatio-temporal model including an in-depth inference analysis for 5-fold cross validation and show the benefits of incorporating sequence data in our framework.

3.1 Model performance

The proposed method achieves an overall per-class recall (mean and standard deviation) of $93.82 \pm 5.11\%$ for detecting an ongoing insertion and $70.88 \pm 11.83\%$ for identifying a reached target press fit for five independent test specimens. Table 1 illustrates the per-class recall and precision for each fold. Hereby, we consider the per-class recall (bold values) as main metric, as it corresponds to the ratio of correctly identified sequences. Furthermore, we present an in-depth analysis of the network performance throughout the whole insertion procedure for each independent test specimen in Figure 3.

Table 1. Performance for 5-fold cross validation

		Spatio-temporal model	Non-sequence baseline	Planned broach size	Specimen number		
Fold 0	recall	94.67%	49.37%	62.35%	53.55%	4	1
	precision	63.96%	90.70%	55.50%	60.49%		
Fold 1	recall	93.50%	69.86%	95.91%	19.31%	4	2
	precision	89.47%	79.69%	74.82%	65.38%		
Fold 2	recall	96.72%	71.74%	42.86%	95.09%	4	3
	precision	81.94%	94.29%	91.58%	57.20%		
Fold 3	recall	99.71%	81.90%	92.08%	38.97%	7	4
	precision	94.23%	98.96%	81.76%	62.35%		
Fold 3	recall	84.51%	81.54%	76.67%	29.33%	7	5
	precision	96.28%	48.18%	85.36%	18.97%		

3.2 Comparison with Non-sequence Data

We compared the proposed spatio-temporal model with a single spectrogram based detection method. Therefore, we implemented the model architecture proposed by Seibold et al. [20], which yielded promising performance for the application in surgical drill breakthrough detection, and evaluate it on single

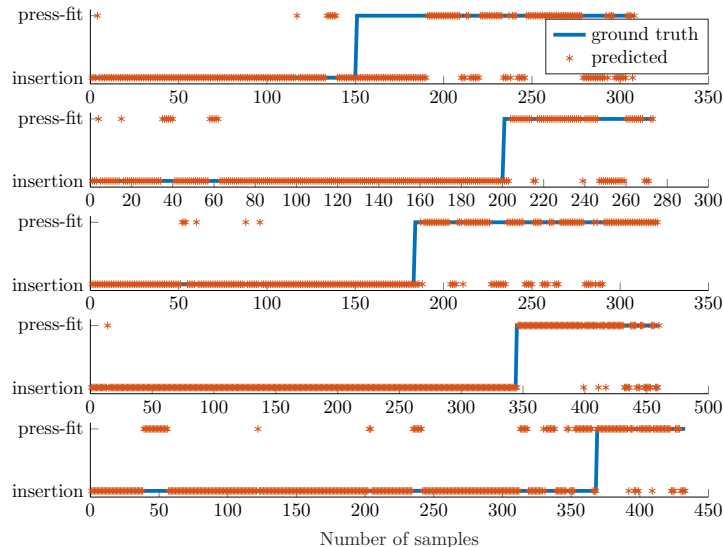


Fig. 3. An in-depth analysis of the results of the spatio-temporal model during 5-fold cross validation. Each plot corresponds to an independent test specimen (Specimen 1-5 in ascending order from top to bottom), the network was trained on the remaining four folds.

spectrograms from the data set collected in the cadaveric experiments. For pre-processing, we applied the same augmentation strategy to the non-sequence data set and normalize every sample spectrogram. Without the temporal context of the sequence of hammer blows the detection performance decreases dramatically. The model reaches a per-class recall of $73.97 \pm 19.59\%$ and $47.25 \pm 26.45\%$ for c_0 (ongoing insertion) and c_1 (press fit reached), respectively. The per-class recall and precision for each fold is given in Table 1.

4 Discussion

The proposed automated method for the assessment of the optimal insertion endpoint of femoral stem implants could be an important step towards reducing the risk of periprosthetic femoral fracture during cementless THA, which would consequently improve patient safety and treatment quality. To the best knowledge of the authors, this is the first work to employ a sensorized instrument for capturing vibroacoustic signals directly from the operation area and to develop a state-of-the-art learning based method for the automated assessment of the seating of the femoral stem implant. We furthermore evaluated the proposed model in cadaver experiments in which the real intervention was simulated as closely as possible on thawed fresh-frozen cadavers.

In comparison to previous work, we demonstrated feasibility of an automated system for the assessment of the optimal insertion endpoint. Even though the full potential of deep learning models and the advantages over handcrafted features is revealed when large amounts of training data are available [11], we show that the proposed model is able to learn useful information for the majority of the presented test cases even though we have a small data set.

The results of 5-fold cross validation show that the proposed network yields promising overall performance, however the model confuses the samples in the critical region of the optimal insertion endpoint (when changing from "insertion" to "press fit" in Figure 3) for the first and last fold. These outliers and the resulting relatively large standard deviation in cross validation can be attributed to the fact that the sample size is relatively small with five cadaveric specimens also due to the relevant inter-subject variance in bone density. An increased size of data would improve the model's capability for generalization and improve the performance of the algorithm. Nevertheless, we consider our sample size sufficient for a technical feasibility study, because human cadaver experiments are associated with significant cost and ethical considerations. More extensive data collection will be addressed in future work together with additional postprocessing steps, such as majority vote or additional convergence criteria, to improve generalization capabilities. Even though the presented dataset is not highly imbalanced, the influence of class imbalance should be investigated in future work.

In the presented work, we showed that the spatio-temporal model clearly outperforms non-sequence data for the application in the assessment of the femoral stem press fit. However, in future work, other temporal modelling approaches, such as Temporal Convolutional Neural Networks (TCNs), could be employed. Furthermore, the influence of additional mechanisms, such as Self-attention, could be investigated. A technical limitation of the presented work is the subjective decision of the surgeon for the definition of the optimal insertion endpoint. However, this definition of the ground truth is in line with all prior work and a quantitative measurement is infeasible, as e.g. a measurement of the pull out force would require an extensive measurement setup and is out of scope of the presented work.

5 Conclusion

To the best knowledge of the authors, we propose the first automated detection approach for assessing the press fit of femoral stem implants in THA. Our method consists of a sensorized smart instrument and a spatio-temporal model capable of inferring the optimal insertion endpoint. The proposed solution shows the general feasibility of such an automated system and is thoroughly evaluated on human cadaveric data with an in-depth analysis of the model performance in 5-fold cross validation using independent test specimens, achieving a per-class sensitivity of $93.82 \pm 5.11\%$ and $70.88 \pm 11.83\%$ for identifying an ongoing insertion or a reached target press fit, respectively.

The presented system could not only be valuable as supplementary system for navigation systems and robotic applications, but also for error prevention in conventional surgery. Additionally, an automated system for the assessment of the optimal femoral stem insertion point could be employed for surgical skill assessment to define the level of surgical expertise.

Acknowledgment

This work is part of the SURGENT project and was funded by University Medicine Zurich/Hochschulmedizin Zürich. Matthias Seibold and Nassir Navab are partly funded by the Balgrist Foundation in form of the guest professorship at Balgrist University Hospital.

References

1. Abdel, M.P., Houdek, M.T., Watts, C.D., Lewallen, D.G., Berry, D.J.: Epidemiology of periprosthetic femoral fractures in 5417 revision total hip arthroplasties: a 40-year experience. *The Bone & Joint Journal* **98**, 468–474 (2016)
2. Berend, K.R., Lombardi, A.V., Mallory, T.H., Chonko, D.J., Dodds, K.L., Adams, J.B.: Cerclage wires or cables for the management of intraoperative fracture associated with a cementless, tapered femoral prosthesis: results at 2 to 16 years. *Journal of Arthroplasty* **19** (2004)
3. Capello, W.N., Houdek, M.T., Watts, C.D., Lewallen, D.G., Berry, D.J.: Periprosthetic fractures around a cementless hydroxyapatite-coated implant: a new fracture pattern is described. *Clinical Orthopaedics and Related Research* **472**, 604–610 (2014)
4. Dubory, A., Rosi, G., Tijou, A., Lomami, H.A., Flouzat-Lachaniette, C.H., Haiat, G.: A cadaveric validation of a method based on impact analysis to monitor the femoral stem insertion. *Journal of the Mechanical Behavior of Biomedical Materials* **103** (2020)
5. Goossens, Q., Leuridan, S., Roosen, J.: Monitoring of reamer seating using acoustic information. In: Annual meeting of the European Society of Biomechanics (2015)
6. Goossens, Q., Pastrav, L., Roosen, J., Mulier, M., Desmet, W., Vander Sloten, J., Denis, K.: Acoustic analysis to monitor implant seating and early detect fractures in cementless tha: An in vivo study. *Journal of Orthopedic Research* (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
8. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R., Wilson, K.: Cnn architectures for large-scale audio classification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135 (2017)
9. Illanes, A., Boese, A., Maldonado, I., Pashazadeh, A., Schaffler, A., Navab, N., Friebe, M.: Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing. *Scientific Reports* **8** (2018)
10. Le Béguet, P., Canovas, F., Roche, O., Goldschild, M., Batard, J.: Uncemented Femoral Stems for Revision Surgery. Springer (2015)

11. Lin, W., Hasenstab, K., Cunha, G.M., Schwartzman, A.: Comparison of hand-crafted features and convolutional neural networks for liver mr image adequacy assessment. *Scientific Reports* **10** (2020)
12. Lyon, D.A.: The discrete fourier transform, part 4: Spectral leakage. *Journal of Object Technology* **8**(7), 23–34 (2009)
13. McFee, B., Raffel, C., Liang, D., PW Ellis, D., McVicar, M., Battenberg, E., Nieto, O.: *librosa: Audio and music signal analysis in python*. In: 14th python in science conference. pp. 18–25 (2015)
14. Morohashi, I., Iwase, H., Kanda, A., Sato, T., Homma, Y., Mogami, A., Obayashi, O., Kaneko, K.: Acoustic pattern evaluation during cementless hip arthroplasty surgery may be a new method for predicting complications. *SICOT-J* **3** (2017)
15. Oberst, S., Baetz, J., Campbell, G., Lampe, F., Lai, J.C.S., Hoffmann, N., Morlock, M.: Vibro-acoustic and nonlinear analysis of cadavric femoral bone impaction in cavity preparations. *International Journal of Mechanical Sciences* **144**, 739–745 (2018)
16. Ostler, D., Seibold, M., Fuchtmann, J., Samm, N., Feussner, H., Wilhelm, D., Navab, N.: Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions. *International Journal of Computer Assisted Radiology and Surgery* (2020)
17. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.y., Sainath, T.: Deep learning for audio signal processing. *IEEE Journal on Selected Topics in Signal Processing* **14**, 206–219 (2019)
18. Renner, L., Janz, V., Perka, C., Wassilew, G.I.: What do we get from navigation in primary tha? *EFORT Open Reviews* **1**, 205—210 (2016)
19. Ricoli, W., Queiroz, M.C., Guimaraes, R.P., honda, E.K., Polesello, G., Fucs, P.M.: Prevalence and risk factors for intra-operative periprosthetic fractures in one thousand eight hundred and seventy two patients undergoing total hip arthroplasty: a cross-sectional study. *International Orthopaedics* **39**, 1939–1943 (2015)
20. Seibold, M., Maurer, S., Hoch, A., Zingg, P., Farshad, M., Navab, N., Fürnstahl, P.: Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery. *Scientific Reports* **11** (2021)
21. Suehn, T., Pandey, A., Friebe, M., Illanes, A., Boese, A., Lohman, C.: Acoustic sensing of tissue-tool interactions – potential applications in arthroscopic surgery. *Current Directions in Biomedical Engineering* **6** (2020)
22. Tijou, A., Rosi, G., Vayron, R., Lomami, H.A., Hernigou, P., Flouzat-Lachaniette, C.H., Haiat, G.: Monitoring cementless femoral stem insertion by impact analyses: An in vitro study. *Journal of the Mechanical Behavior of Biomedical Materials* **88**, 102–108 (2018)
23. Yun, H.H., Lim, J.T., Yang, S.H., Park, P.S.: Occult periprosthetic femoral fractures occur frequently during a long, trapezoidal, double-tapered cementless femoral stem fixation in primary tha. *PLoS One* **19** (2019)

3.4 Approaches based on Raw Waveforms

In spectrogram computation, fixed filters, which can be spaced linearly or logarithmically on the frequency scale or mapped to a non-linear scale like the Mel scale, are applied to raw audio waveforms. Instead of predefining the filter bank as a fixed preprocessing step, the filters can also be jointly learned in the training process which can yield superior results, e.g. for large-scale environmental audio classification tasks [24]. Furthermore, it is possible to directly operate on raw audio signals, e.g. using dilated causal convolutions which have a large 1D receptive field and therefore can deal with long-range temporal dependencies as present in acoustic signals [77].

While raw audio-based models have achieved impressive sound quality in synthesis tasks, where the human-perceived quality of the generated audio is crucial and spectrogram-based methods fall back because phase information is lost in spectrogram generation, raw waveforms have not yet prevailed as a representation for learning based audio analysis tasks. Raw audio imposes higher computational cost and data requirements compared to spectrogram input. Therefore, mel spectrogram-based systems achieve state-of-the-art results in practical applications of acoustic sensing. [87]

3.5 Data Augmentation

For the application in audio-related tasks such as speech processing, music, or environmental audio classification, large-scale datasets have been established, e.g. the Librispeech [80], UrbanSound-8K [95], or Youtube-8M [2] datasets. In contrast to these large datasets, which are often collected using internet resources, data collection in a realistic clinical environment is very expensive and even impossible in comparable scales. Therefore, the development of clinical acoustic sensing systems introduces the individual challenge of training with sparse, often imbalanced datasets.

A way to handle sparse data is to artificially increase the corpus of training data using data augmentation methods. For audio signals, a variety of established data augmentation methods have been proposed in previous work [118]. There are two main differences for audio augmentation methods as they either work with raw waveform data, such as adding noise, pitch shifting, time stretching, etc., or are applied to audio spectrograms directly, such as SpecAugment [81]. First applications of using generative neural networks for the augmentation of audio data, e.g. for speech [13] or environmental audio data [61], have been proposed in the literature.

3.5.1 Contribution: Conditional Generative Data Augmentation for Clinical Audio Datasets (MICCAI 2022)

Summary: In the clinical context, the creation of large-scale datasets requires access to the realistic environment, e.g. the operating room, and medical regulations impose strict rules

on the usage of personalized data. As deep learning-based systems require large amounts of data for training and good generalization, data augmentation is an essential tool to artificially enlarge the amount of training data and stabilize the training process. For the augmentation of audio data, classical data augmentation methods are either applied to raw audio, such as adding noise, time stretching or pitch shifting the signal, or spectrogram-based, such as time and frequency masking applied directly to spectrograms. These methods, however, do not necessarily generate samples that could be captured in a real environment. In this work, we propose a method based on a conditional generative adversarial network to synthesize samples from the learned data distribution of a dataset. We evaluate the proposed augmentation method on a dataset captured in the operating room which contains sound samples of typical surgical actions in Total Hip Arthroplasty. The method's ability of generating realistic class-conditioned samples is shown and the quality of the augmentations is evaluated in terms of classification performance of a ResNet-18 classifier trained on the proposed dataset. We compare the method with classical and established augmentation methods and show that the proposed method achieves superior results. The presented augmentation method has therefore great potential to improve the data bottleneck for the training of medical acoustic sensing systems.

Contributions: The author of this thesis was responsible for formulating the problem and approach with medical consultations from Armando Hoch and Mazda Farshad as well as for capturing the data in the operating room, designing the data processing, detection, and evaluation pipeline and writing the manuscript. Nassir Navab and Philipp Fürnstahl contributed in the form of discussions and feedback throughout the whole project and for proofreading of the manuscript.

Copyright Statement: Reprint of the final author's accepted manuscript by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2022, Lecture Notes in Computer Science, vol 13437, Matthias Seibold, Armando Hoch, Mazda Farshad, Nassir Navab & Philipp Fürnstahl, "Conditional Generative Data Augmentation for Clinical Audio Datasets" (2022).

The published version of this work can be accessed under: https://doi.org/10.1007/978-3-031-16449-1_33

Conditional Generative Data Augmentation for Clinical Audio Datasets

Matthias Seibold^{1,2}, Armando Hoch³, Mazda Farshad³, *Nassir Navab¹, and
*Philipp Fürnstahl^{2,3}

¹ Computer Aided Medical Procedures (CAMP), Technical University of Munich,
Munich, DE-85748, Germany

² Research in Orthopedic Computer Science (ROCS), University Hospital Balgrist,
University of Zurich, Zurich, CH-8008, Switzerland

³ Balgrist University Hospital, Zurich, CH-8008, Switzerland
*equally contributing last authors

Abstract. In this work, we propose a novel data augmentation method for clinical audio datasets based on a conditional Wasserstein Generative Adversarial Network with Gradient Penalty (cWGAN-GP), operating on log-mel spectrograms. To validate our method, we created a clinical audio dataset which was recorded in a real-world operating room during Total Hip Arthroplasty (THA) procedures and contains typical sounds which resemble the different phases of the intervention. We demonstrate the capability of the proposed method to generate realistic class-conditioned samples from the dataset distribution and show that training with the generated augmented samples outperforms classical audio augmentation methods in terms of classification performance. The performance was evaluated using a ResNet-18 classifier which shows a mean Macro F1-score improvement of 1.70% in a 5-fold cross validation experiment using the proposed augmentation method. Because clinical data is often expensive to acquire, the development of realistic and high-quality data augmentation methods is crucial to improve the robustness and generalization capabilities of learning-based algorithms which is especially important for safety-critical medical applications. Therefore, the proposed data augmentation method is an important step towards improving the data bottleneck for clinical audio-based machine learning systems.

Keywords: Deep Learning · Data Augmentation · Acoustic Sensing · Total Hip Arthroplasty · Generative Adversarial Networks

1 Introduction

Acoustic signals are easy and low-cost to acquire, can be captured using airborne or contact microphones and show great potentials in medical applications for interventional guidance and support systems. Successful applications are intra-operative tissue characterization during needle insertion [9] and tissue coagulation [16], the identification of the insertion endpoint in THA procedures

[3,20], error prevention in surgical drilling during orthopedic procedures [21] or guidance in orthopedic arthroscopy procedures [23].

Furthermore, acoustic signals have successfully been employed for diagnostic medical applications. Exemplary applications include the assessment of cartilage degeneration by measuring structure borne noise in the human knee during movement [11], the development of a prototype for the detection of implant loosening through an acoustic sensor system [2], a system for monitoring the acoustic emissions of THA implants [19], or the automated analysis of lung sounds captured with a digital stethoscope which allows non-specialists to screen for pulmonary fibrosis [14].

Through recent advances in machine learning research, learning-based methods have replaced and outperformed classical acoustic signal processing-based approaches, as well as classical handcrafted feature-based learning approaches for many acoustic audio signal processing tasks [18]. However, state-of-the-art deep learning methods require large amounts of training data to achieve superior performance and generalize well to unseen data, which are often difficult or infeasible to acquire in a clinical setting. To tackle this issue, the usage of augmentation techniques is a standard approach to increase the diversity and size of training datasets. Hereby, new samples can be synthesized by applying transformations to the existing data, e.g. rotation and cropping for images, replacing words with synonyms for text, and applying noise, pitch shifting, and time stretching to audio samples [28]. Even though these data augmentation methods improve the performance of target applications, they do not necessarily generate realistic samples which is especially crucial in the medical domain where reliability is a key factor. One solution to this problem is for example to exploit the underlying physics for augmentation, e.g. for ultrasound image augmentation [26] which is, however, not applicable for clinical audio data. In the presented work, we will focus on realistic data augmentation of audio datasets for medical applications.

Recently, deep generative models, a family of deep learning models, which are able to synthesize realistic samples from a learned distribution, have been applied for data augmentation of various data modalities outside of the medical domain. For the augmentation of audio data, different generative approaches have been introduced, of which related work to the proposed method is described in the following section. Hu et al. utilized a GAN to synthesize samples of logarithmic Mel-filter bank coefficients (FBANK) from a learned distribution of a speech dataset and subsequently generated soft labels using a pretrained classifier [8]. Madhu et al. trained separate GANs on mel-spectrograms for each class of a dataset to generate augmentation data [12]. Chatziagapi et al. used the Balancing GAN (BAGAN) framework [13] to augment an imbalanced speech dataset [1]. A conditional GAN was employed for data augmentation of speech using FBANK features by Sheng et al. [22] and for respiratory audio signals based on raw waveform augmentation by Jayalakshmy et al. [10].

In this work, we introduce a novel augmentation technique for audio data based on a conditional Wasserstein GAN model with Gradient Penalty (cWGAN-

GP) which produces higher-quality samples and is easier to train than standard GANs [5]. The proposed model operates on log-mel spectrograms which have been shown to outperform other feature representations and achieves state-of-the-art performance in audio classification tasks [18]. The proposed model is able to generate realistic and high-quality log-mel spectrograms from the learned dataset distribution. We show that our model can be used for two augmentation strategies, doubling the number of samples and balancing the dataset. While classical audio augmentation techniques might improve the performance of the classifier, they do not generate samples that can be captured in a real environment and might therefore be inconsistent with the real variability of captured real-world acoustic signals. In contrast, the proposed model is able to generate realistic samples from the learned distribution of the original data.

To evaluate the proposed framework on realistic clinical data, we introduce a novel audio dataset containing sounds of surgical actions recorded from five real THA procedures which resemble the different phases of the intervention. We thoroughly evaluate the proposed method on the proposed dataset in terms of classification performance improvement of a ResNet-18 classifier with and without data augmentation using 5-fold cross validation and compare the results with classical audio augmentation techniques.

2 Materials and Method

2.1 Novel Surgical Audio Dataset

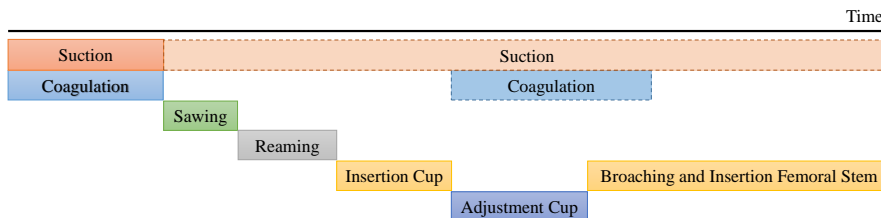


Fig. 1. The classes of the novel clinical dataset resemble the phases of a THA procedure. Occurrences with drawn through lines indicate intensive usage of the respective surgical action, dashed lines correspond to sporadic usage.

Figure 1 illustrates the occurrence of the six classes $C := \{\text{Suction, Coagulation, Sawing, Reaming, Insertion, Adjustment}\}$ present in the dataset over the course of a THA procedure. Please note that "Insertion Cup" and "Broaching and Insertion Femoral Stem" were joined into a single class ("Insertion") because of the similar acoustic signature generated by hammering onto the metal structure of the insertion tools for the acetabular cup, femoral broach and femoral stem implant, respectively. The "Adjustment" class also contains

hammering signals that are, however, performed with a screwdriver-like tool which is used to adjust the orientation of the acetabular cup and generates a slightly different sound. During opening the access to the area of operation in the beginning of the procedure, suction and coagulation is employed intensively, whereas in the rest of the procedure both surgical actions are performed sporadically and on demand (indicated through dashed outlines in Figure 1). All samples were manually cut from recordings of five THA interventions conducted at our university hospital for which we captured audio with a framerate of 44.1 kHz using a air-borne shotgun microphone (Røde NTG2) pointed towards the area of operation and video captured from the OR light camera (Trumpf TruVidia). The captured video was used to facilitate the labelling process. We labelled the dataset in a way that audio samples do not contain overlapping classes and no staff conversations. An ethical approval has been obtained prior to recording the data in the operating room. The resulting dataset contains 568 recordings with a length of 1 s to 31 s and the following distribution: $n_{raw,Adjustment} = 68$, $n_{raw,Coagulation} = 117$, $n_{raw,Insertion} = 76$, $n_{raw,Reaming} = 64$, $n_{raw,Sawing} = 21$, and $n_{raw,Suction} = 222$. The dataset can be accessed under <https://rocs.balgrist.ch/en/open-access/>.

2.2 Data Preprocessing and Baseline Augmentations

Log-mel spectrograms are a two-dimensional representation of an audio signal, mapping frequency components of a signal to the ordinate and time to the abscissa. They offer a dense representation of the signal, reduce the dimensionality of the samples, and have been shown to yield superior classification performance for a wide variety of acoustic sensing tasks [18]. We compute log-mel spectrograms of size 64×64 from the dataset samples by applying a sliding window technique with non-overlapping windows of length $L = 16380$ samples, a Short Time Fourier Transform (STFT) hop length of $H = 256$ samples and $n_{mels} = 64$ mel bins using the Python library *librosa 0.8.1* [15]. We compute a total number of 3597 individual spectrograms from the raw waveform dataset. The resulting number of spectrograms per-class is: $n_{spec,Adjustment} = 494$, $n_{spec,Coagulation} = 608$, $n_{spec,Insertion} = 967$, $n_{spec,Reaming} = 469$, $n_{spec,Sawing} = 160$, and $n_{spec,Suction} = 899$. For the evaluation using 5-fold cross validation, we randomly split the dataset into five folds on the raw waveform level over all recordings, as the recording conditions are identical.

To compare the proposed augmentation method against classical signal processing augmentation approaches, we implemented the following augmentation strategies which are applied to the raw waveforms directly. We apply Gaussian noise with $\mu = 0$ and $\sigma = 0.01$. We apply Pitch Shifting by 3 semitones upwards. We apply time stretching with a factor of 1.5. Furthermore, we compare our method with *SpecAugment*, a widely used approach for audio augmentation in Automatic Speech Recognition (ASR) tasks which applies random time-warping, frequency- and time-masking to the spectrograms directly [17]. For a fair comparison of all augmentations, we add 100% generated samples for

each augmentation strategy, respectively. We normalize the data by computing $X_{norm,mel} = (X_{mel} - \mu)/\sigma$, where (μ) is the mean and (σ) is the standard deviation of the entire dataset.

2.3 Conditional Generative Data Augmentation Method

The architectural details of the proposed GAN are illustrated in Figure 2. To stabilize the training process and improve the generated sample quality, we apply the Wasserstein loss with Gradient Penalty (GP) as introduced by Gulrajani et al. [5] which enforces a constraint such that the gradients of the discriminator’s (critic) output w.r.t the inputs have unit norm. This approach greatly improves the stability of the training and compensates for problems such as mode collapse. We define the critic’s loss function as:

$$L_C = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x}, y)] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x, y)] + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [\|\Delta_{\hat{x}} D(\hat{x}, y)\|_2 - 1]^2 \quad (1)$$

where \mathbb{P}_r is the real distribution, \mathbb{P}_g is the generated distribution, and $\mathbb{P}_{\hat{x}}$ is the interpolated distribution. The interpolated samples \hat{x} are uniformly sampled along a straight line between real x and generated \tilde{x} samples by computing:

$$\hat{x} = \epsilon x + (1 - \epsilon)\tilde{x} \quad (2)$$

We use the recommended GP weight of $\lambda = 10$, a batch size of 64 and train the discriminator five times for each generator iteration. In order to choose the stopping point for training, we frequently compute the Fréchet Inception Distance (FID) [7] which is calculated from features of a pretrained classifier by:

$$FID = \|\mu_r - \mu_g\|^2 + \text{Tr}(C_r + C_g - 2 * \sqrt{C_r * C_g}) \quad (3)$$

as a measure of the quality for the generated samples and stop the training at epoch 580. Hereby, μ_r and μ_g represent the feature-wise mean of the real and generated spectrograms, C_r and C_g the respective covariance matrices. Because of the structural differences of images and spectrograms, we cannot use an Inception v3 network pretrained on ImageNet to compute the FID. Therefore, we employ a ResNet-18 [6] model pretrained on the proposed dataset, extract the features from the last convolutional layer, and use these features for FID calculation. The proposed model is implemented with *TensorFlow/Keras 2.6* and trained using the Adam optimizer ($LR = 1e - 4$, $\beta_1 = 0.5$, $\beta_2 = 0.9$) in ~ 6 hours on a NVidia RTX 2080 SUPER GPU.

A nonlinear activation function is omitted in the last convolutional layer of the generator because the spectrogram samples are not normalized in the range $[0, 1]$. The mapping layer of the generator employs a dense layer, whereas in the discriminator (critic) we use repeat and reshaping operations for remapping. The generator and discriminator have a total number of 1,526,084 and 4,321,153 parameters, respectively. The implementation, pretrained models, and dataset can be accessed under: <https://rocs.balgrist.ch/en/open-access/>.

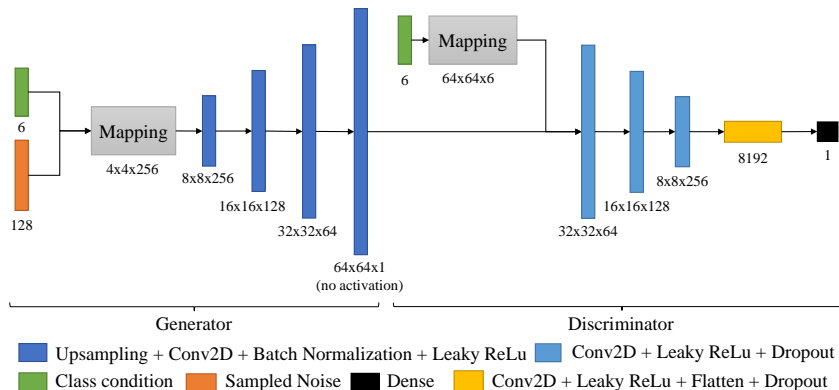


Fig. 2. The architecture of the proposed model including output sizes of each layer. The input for the generator is a noise vector of size 1x128 and a class condition. The generator outputs a spectrogram which is fed to the discriminator together with the class condition. The discriminator (critic) outputs a scalar realness score.

2.4 Classification Model

To evaluate the augmentation performance of our model against classical audio augmentation techniques, we analyze the effect of augmentation on the classification performance in a 5-fold cross validation experiment using a ResNet-18 [6] classifier, an architecture which has been successfully employed for clinical audio classification tasks [20,21]. We train the classifier from scratch for 20 epochs using categorical crossentropy loss, the Adam optimizer ($LR = 1e - 4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and a batch size of 32.

3 Results and Evaluation

In Figure 3, we show a comparison of randomly chosen ground truth and randomly generated samples. The visual quality of the generated samples is comparable to the original data and the model seems to be able to generate samples conditioned on the queried class. By further visual inspection it can be observed that the synthesized samples contain the characteristics of the original dataset, e.g. the hammer strokes are clearly visible for the classes "Adjustment" and "Insertion".

The quantitative evaluation of the classification performance using a ResNet-18 classifier is given in Table 1. We report the mean Macro F1-Score in the format $mean \pm std.$. We compare training without augmentations and classical audio augmentation techniques (adding noise, pitch shifting, time stretching, and SpecAugment [17]) with the proposed method. The cWGAN-GP-based augmentations outperform all classical augmentation strategies when doubling the samples (+1.70%) and show similar performance (+1.07%) as the best performing classical augmentation strategy (Time Stretch) when balancing the dataset.

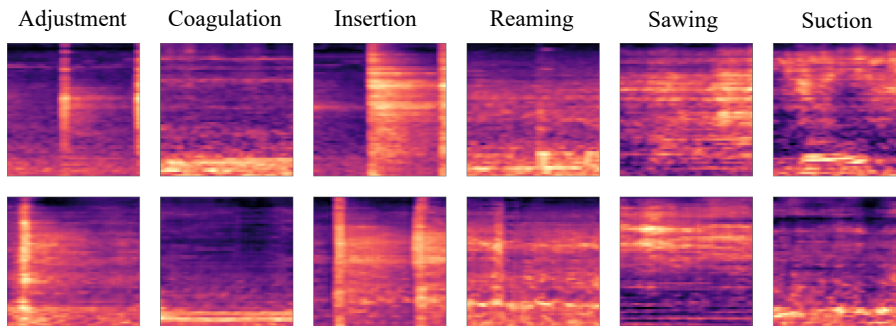


Fig. 3. The top row shows log-mel spectrograms of random samples for each class present in the acquired dataset, the bottom row shows log-mel spectrograms generated by the proposed model for each class, respectively.

Augmentation Technique	Mean Macro F1-Score	Relative Improvement
No Augmentation	$93.90 \pm 2.48\%$	
Add Noise	$92.87 \pm 0.99\%$	-1.03%
Pitch Shift	$94.73 \pm 1.28\%$	+0.83%
Time Stretch	$95.00 \pm 1.49\%$	+1.10%
SpecAugment [17]	$94.23 \pm 1.14\%$	+0.33%
cWGAN-GP (balanced)	$94.97 \pm 1.71\%$	+1.07%
cWGAN-GP (doubled)	$95.60 \pm 1.26\%$	+1.70%

Table 1. Results of the proposed model in comparison to classical audio augmentation techniques.

4 Discussion

The proposed augmentation method is an important step towards improving the data limitations by generating synthetic in-distribution augmentation data for clinical applications for which it is often expensive or even impossible to gather large amounts of training data. We showed that our augmentation strategy outperforms classical signal processing approaches and has the capability to balance imbalanced datasets to a certain extent. To balance imbalanced datasets, any arbitrary number of samples can easily be generated for each class with the proposed approach which is not possible using classical signal processing techniques in the same way. However, for the given dataset and configuration, doubling the number of samples using the proposed augmentation method leads to the best final classification results. Furthermore, we show that the proposed method outperforms SpecAugment [17], an established audio augmentation method which applies time-warping, as well as frequency and time masking to the spectrogram data directly.

In future work, we want to benchmark the proposed framework with other generative augmentation models and model architectures, investigate the perfor-

mance of the proposed approach on more (balanced and imbalanced) datasets and further optimize our model towards improved classification performance. Furthermore, it should be investigated how combinations of augmentation techniques influence the resulting classification performance and if it is possible to maximize the impact of augmentations through an optimized combination scheme.

The improved performance achieved by the proposed augmentation method comes at the cost of increased demands on computational power and resources. While signal processing augmentations are computed in the range of seconds to minutes, our model requires an additional training step which takes ~ 6 hours for the presented dataset and increases with larger datasets.

Because we created the proposed clinical dataset in a way that it resembles the phases and surgical actions executed during a real THA procedure, potential future clinical applications are the prediction of surgical actions from captured audio signals in the operating room which could be used for workflow recognition and surgical phase detection. Therefore, we consider the proposed dataset as an important step towards automated audio-based clinical workflow detection systems, a topic which has only been studied rudimentally so far [24,27]. The proposed approach is designed to work with spectrogram based audio, which can be transformed back to the signal domain, e.g. using the Griffin-Lim algorithm [4] or more recently introduced learning-based transformation approaches, e.g. the work by Takamichi et al. [25]. We reconstructed waveforms from a few generated spectrograms using the Griffin-Lim algorithm and could, despite artifacts being present, recognize acoustic similarities to the original samples for each class, respectively. In future work, the proposed augmentation method could furthermore be transferred to other medical and non-medical grid-like data domains.

5 Conclusion

In the presented work, we introduce a novel data augmentation method for medical audio data and evaluate it on a clinical dataset which was recorded in real-world Total Hip Arthroplasty (THA) surgeries. The proposed dataset contains sound samples of six surgical actions which resemble the different phases of a THA intervention. We show in quantitative evaluations that the proposed method outperforms classical signal and spectrogram processing-based augmentation techniques in terms of Mean Macro F1-Score, evaluated using a ResNet-18 classifier in a 5-fold cross validation experiment. By generating high-quality in-distribution samples for data augmentation, our method has the potential to improve the data bottleneck for acoustic learning-based medical support systems.

Acknowledgment

This work is part of the SURGENT project under the umbrella of Hochschulmedizin Zürich.

References

1. Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., Narayanan, S.: Data Augmentation Using GANs for Speech Emotion Recognition. In: Proc. Interspeech 2019. pp. 171–175 (2019)
2. Ewald, H., Timm, U., Ruther, C., Mittelmeier, W., Bader, R., Kluess, D.: Acoustic sensor system for loosening detection of hip implants. In: 2011 Fifth International Conference on Sensing Technology. pp. 494–497 (2011)
3. Goossens, Q., Pastrav, L., Roosen, J., Mulier, M., Desmet, W., Vander Sloten, J., Denis, K.: Acoustic analysis to monitor implant seating and early detect fractures in cementless tha: An in vivo study. *Journal of Orthopedic Research* (2020)
4. Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **32**(2), 236–243 (1984)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5769–5779 (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640 (2017)
8. Hu, H., Tan, T., Qian, Y.: Generative adversarial networks based data augmentation for noise robust speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5044–5048 (2018)
9. Illanes, A., Boese, A., Maldonado, I., Pashazadeh, A., Schaufler, A., Navab, N., Friebe, M.: Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing. *Scientific Reports* **8** (2018)
10. Jayalakshmy, S., Sudha, G.F.: Conditional gan based augmentation for predictive modeling of respiratory signals. *Computers in Biology and Medicine* **138**, 104930 (2021)
11. Kim, K.S., Seo, J.H., Kang, J.U., Song, C.G.: An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis. *Computer Methods and Programs in Biomedicine* **94**(2), 198–206 (2009)
12. Madhu, A., Kumaraswamy, S.: Data augmentation using generative adversarial network for environmental sound classification. In: 2019 27th European Signal Processing Conference (EUSIPCO) (2019)
13. Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, A.C.I.: Bagan: Data augmentation with balancing gan. *arXiv abs/1803.09655* (2018)
14. Marshall, A., Boussakta, S.: Signal analysis of medical acoustic sounds with applications to chest medicine. *Journal of the Franklin Institute* **344**(3), 230–242 (2007)
15. McFee, B., Raffel, C., Liang, D., PW Ellis, D., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: 14th python in science conference. pp. 18–25 (2015)

16. Ostler, D., Seibold, M., Fuchtmann, J., Sann, N., Feussner, H., Wilhelm, D., Navab, N.: Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions. *International Journal of Computer Assisted Radiology and Surgery* (2020)
17. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: SpecAugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019* (Sep 2019)
18. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.y., Sainath, T.: Deep learning for audio signal processing. *IEEE Journal on Selected Topics in Signal Processing* **14**, 206–219 (2019)
19. Rodgers, G.W., Young, J.L., Fields, A.V., Shearer, R.Z., Woodfield, T.B.F., Hooper, G.J., Chase, J.G.: Acoustic emission monitoring of total hip arthroplasty implants. *IFAC Proceedings Volumes* **47**(3), 4796–4800 (2014), 19th IFAC World Congress
20. Seibold, M., Hoch, A., Suter, D., Farshad, M., Zingg, P.O., Navab, N., Fürnstahl, P.: Acoustic-based spatio-temporal learning for press-fit evaluation of femoral stem implants. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 447–456 (2021)
21. Seibold, M., Maurer, S., Hoch, A., Zingg, P., Farshad, M., Navab, N., Fürnstahl, P.: Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery. *Scientific Reports* **11** (2021)
22. Sheng, P., Yang, Z., Hu, H., Tan, T., Qian, Y.: Data augmentation using conditional generative adversarial networks for robust speech recognition. In: *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. pp. 121–125 (2018)
23. Suehn, T., Pandey, A., Friebe, M., Illanes, A., Boese, A., Lohman, C.: Acoustic sensing of tissue-tool interactions – potential applications in arthroscopic surgery. *Current Directions in Biomedical Engineering* **6** (2020)
24. Suzuki, T., Sakurai, Y., Yoshimitsu, K., Nambu, K., Muragaki, Y., Iseki, H.: Intraoperative multichannel audio-visual information recording and automatic surgical phase and incident detection. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. pp. 1190–1193 (2010)
25. Takamichi, S., Saito, Y., Takamune, N., Kitamura, D., Saruwatari, H.: Phase reconstruction from amplitude spectrograms based on directional-statistics deep neural networks. *Signal Processing* **169**, 107368 (2020)
26. Tirindelli, M., Eilers, C., Simson, W., Paschali, M., Azampour, M.F., Navab, N.: Rethinking ultrasound augmentation: A physics-inspired approach. In: *Medical Image Computing and Computer Assisted Intervention*. pp. 690–700 (2021)
27. Weede, O., Dittrich, F., Wörn, H., Jensen, B., Knoll, A., Wilhelm, D., Krantzfelder, M., Schneider, A., Feussner, H.: Workflow analysis and surgical phase recognition in minimally invasive surgery. In: *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. pp. 1080–1074 (2012)
28. Wei, S., Zou, S., Liao, F., Lang, W.: A comparison on data augmentation methods based on deep learning for audio classification. *Journal of Physics: Conference Series* **1453**(1), 012085 (2020)


3.5.2 Contribution: Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data (MICAD 2022)

Summary: As described in the previous publication, data augmentation is an essential tool for the design of deep learning systems to improve the stability of the training and artificially diversify the training data. As the previously proposed method based on a conditional generative adversarial network has been shown to be a promising approach to sample novel synthetic data points from a learned data set distribution and improve the performance of audio analysis downstream tasks, we continued the project and improved the architecture of the proposed model in the work presented below. Therefore, we integrated residual Squeeze and Excitation modules in the generator of the two-component network and showed that the proposed modification results in a reduced redundancy in the latent features of the model through the analysis of feature correlation. We could improve the Macro F1-Score of a classifier trained on the previously proposed THA data set by 1.14% in comparison to the previous method which corresponds to a performance improvement of 2.84% in regard to training without augmentations. The presented work is a further contribution and an important step towards improving the problem of data limitations for the design of medical deep learning-based acoustic sensing systems.

Contributions: The author of this thesis was responsible for formulating the problem and approach, conceptualizing the paper, and writing the manuscript together with Mane Margaryan. Indu Joshi gave technical advice and feedback on the work and supported the writing process. Mazda Farshad, Nassir Navab and Philipp Frnstahl contributed in the form of discussions and feedback throughout the whole project and for proofreading of the manuscript.

Copyright Statement: This work has been accepted for the The 3rd International Conference on Medical Imaging and Computer-Aided Diagnosis 2022 in Leicester, UK, taking place on Nov. 20-21, 2022. The preprint has been published under <https://arxiv.org/abs/2210.16170>


Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data

Mane Margaryan^{1,*}, Matthias Seibold^{1,2,*}, Indu Joshi¹, Mazda Farshad³,
Philipp Fürnstahl², Nassir Navab¹

¹ Computer Aided Medical Procedures, Technical University Munich, Germany

² Research in Orthopedic Computer Science, Balgrist University Hospital, University of Zurich, Switzerland

³ Department of Orthopedics, Balgrist University Hospital, University of Zurich, Switzerland

 matthias.seibold@balgrist.ch

* equally contributing first authors in alphabetical order

Abstract. Data augmentation is a valuable tool for the design of deep learning systems to overcome data limitations and stabilize the training process. Especially in the medical domain, where the collection of large-scale data sets is challenging and expensive due to limited access to patient data, relevant environments, as well as strict regulations, community-curated large-scale public datasets, pretrained models, and advanced data augmentation methods are the main factors for developing reliable systems to improve patient care. However, for the development of medical acoustic sensing systems, an emerging field of research, the community lacks large-scale publicly available data sets and pretrained models. To address the problem of limited data, we propose a conditional generative adversarial neural network-based augmentation method which is able to synthesize mel spectrograms from a learned data distribution of a source data set. In contrast to previously proposed fully convolutional models, the proposed model implements residual Squeeze and Excitation modules in the generator architecture. We show that our method outperforms all classical audio augmentation techniques and previously published generative methods in terms of generated sample quality and a performance improvement of 2.84% of Macro F1-Score for a classifier trained on the augmented data set, an enhancement of 1.14% in relation to previous work. By analyzing the correlation of intermediate feature spaces, we show that the residual Squeeze and Excitation modules help the model to reduce redundancy in the latent features. Therefore, the proposed model advances the state-of-the-art in the augmentation of clinical audio data and improves the data bottleneck for the design of clinical acoustic sensing systems.

Keywords: Generative Neural Networks, Data Augmentation, Audio Signal Processing, Acoustic Sensing, Computer Aided Medicine

1 Introduction

Medical acoustic sensing systems utilize air- and structure-borne acoustic signals that can be captured in a medical environment, such as vibration signals from surgical tools captured with contact microphones [1] or sounds acquired with air-borne microphones directly from the area of operation [2], to provide guidance and support in medical interventions and diagnostics. Because acoustic signals can be captured non-invasively and radiation-free, and the systems are low-cost and easy-to-integrate, acoustic sensing has great potential for the design of multimodal sensing paradigms for the support of human surgeons, surgical diagnostics, robotic surgery, or to analyze surgical workflow. Hereby, acoustic sensing can be used to obtain measurements for applications where conventional medical computer aided support systems are limited, for example for the assessment of implant-bone press-fit which is impossible to obtain using imaging or navigation [1,2] or to complement the limitations of medical imaging for the assessment of implant loosening [3] or cartilage degeneration [4].

Exemplary applications for the successful application of acoustic sensing in medical interventions are error prevention in orthopedic surgery by analyzing drill vibrations to detect drill breakthrough [5], the evaluation of implant seating during insertion of the femoral stem component in Total Hip Arthroplasty (THA) [2,1], or the guidance of the insertion process of surgical needles using structure-borne acoustic signals acquired from the distal end of the medical device [6]. Also in medical diagnostics, acoustic signals have been successfully employed, e.g. for cough detection [7] or the examination of heart sounds [8].

In the recent years, deep learning-based analysis methods have outperformed classical signal processing and machine learning techniques for the processing of acoustic signals [9] which has also been applied in the medical domain in first use cases and showed promising performance improvements [1,5]. While these methods are very powerful, they require large-scale high-quality training data to achieve superior performance and generalization to unseen cases. One of the main challenges for medical applications, however, is the limited availability of large amounts of data due to the limited access to the real surgical environment, expensive acquisition of realistic data, and clinical requirements and regulations. While in the non-medical domain of audio deep learning research, large-scale audio datasets, such as the Librispeech dataset for speech recordings [10] or the UrbanSound-8K dataset for environmental audio [11], are publicly available, the medical domain is lacking large-scale community data for the development of medical acoustic sensing systems. Therefore, especially in the medical domain, data augmentation is a valuable tool to artificially increase the size of a training data set to increase the diversity of training examples and stabilize the training process. To address this issue, we published a medical audio dataset in a previous work which contains acoustic signals recorded in the real operating room during THA procedures which resemble typical surgical actions such as hammering, drilling, or sawing [12] and proposed a data augmentation method based on a conditional generative adversarial network.

However, we note that several studies report that deep networks tend to learn redundant features due to the huge model capacity [13] [14] [15]. Channel attention has been successfully exploited to model channel level dependencies and facilitate learning of less redundant features [16] [17] [18] and subsequently improved model performance. Motivated by these observations, in this paper, we demonstrate that due to the huge number of model parameters, conditional generative adversarial network (cWGAN-GP [12]) learns redundant features. To combat this, we introduce a channel-wise attention mechanism in the generator sub-network through the implementation of Squeeze & Excitation [16] block and residual skip connections [19]. We provide visualizations that signify the reduced redundancy and subsequently, improved quality of generated mel spectrograms samples quantified by a custom version of the Fréchet Inception Distance [20]. As a result, the present work advances the state-of-the-art in data augmentation for the emerging field of medical acoustic sensing and addresses the important issue of data limitations for medical deep learning-based systems.

2 Materials and Methods

2.1 Data set, Preprocessing, and Benchmark Augmentations

We use a publicly available data set⁴ [12] recorded during real Total Hip Arthroplasty surgeries and contains sounds of the typical surgical actions that are performed during the intervention and roughly resemble the different phases of the procedure. The data set includes 568 recordings with a length of 1 s to 31 s and the following distribution: $n_{raw,Adjustment} = 68$, $n_{raw,Coagulation} = 117$, $n_{raw,Insertion} = 76$, $n_{raw,Reaming} = 64$, $n_{raw,Sawing} = 21$, and $n_{raw,Suction} = 222$.

We compute mel spectrograms, a feature representation for audio signals that obtains state-of-the-art results for deep learning-based audio signal processing systems [9], using non-overlapping sliding windows which results in the following sample distribution for the entire data set: $n_{spec,Adjustment} = 494$, $n_{spec,Coagulation} = 608$, $n_{spec,Insertion} = 967$, $n_{spec,Reaming} = 469$, $n_{spec,Sawing} = 160$, and $n_{spec,Suction} = 899$. Mel spectrograms provide a compact representation, capture time- and frequency-domain aspects about a signal and can be computed from a raw waveform by first computing the Short-time Fourier Transform (STFT) X and then filtering the resulting spectra using a triangular filter bank spaced evenly on the mel scale [21] to compute the mel spectrogram X_{mel} . All spectrograms computed within the present work have dimensions 64×64 and are normalized using the formula $X_{norm} = (X_{mel} - \mu)/\sigma$ where μ is the mean and σ is the standard deviation computed over the entire data set.

A number of data augmentation techniques for acoustic signals have been proposed in prior research, among them classical raw signal based methods like adding noise, time stretching, and pitch shifting, as well as spectrogram-based methods, e.g. SpecAugment [22]. Furthermore, we compare the results of the

⁴ The data set can be obtained from: <https://rocs.balgrist.ch/open-access/>

proposed data augmentation framework with the results reported in our previous work [12] in which a standard convolutional conditional generative adversarial network with Wasserstein Loss with Gradient Penalty regularization [23] was employed.

2.2 Proposed Data Augmentation Method

The architecture of the proposed GAN’s generator is depicted in the Figure 1. It consists of 4 convolutional upsampling blocks followed by a squeeze-and-excitation block with a residual connection, a technique originally proposed in by Hu et al. [16]. The Squeeze and Excitation block consists of a global average pooling layer, which allows to *squeeze* global information to channel descriptors, a re-calibration part, which acts as a channel-wise attention mechanism and allows to capture channel-wise relationships in a non-mutually-exclusive way. The last operation scales the input’s channels by multiplying them with the obtained coefficients. The Squeeze and Excitation mechanisms adds two fully connected layers with a ReLU activation function in between and a sigmoid function applied in the end as shown in the equation 1.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \quad (1)$$

Here the variables W_1 and W_2 have the dimensions $(\frac{C}{r} \times C)$ and $(C \times \frac{C}{r})$, respectively, σ is the sigmoid function and δ refers to a ReLU activation. The value of r is a hyperparameter and for our method it was chosen equal to 16 in an empirical manner.

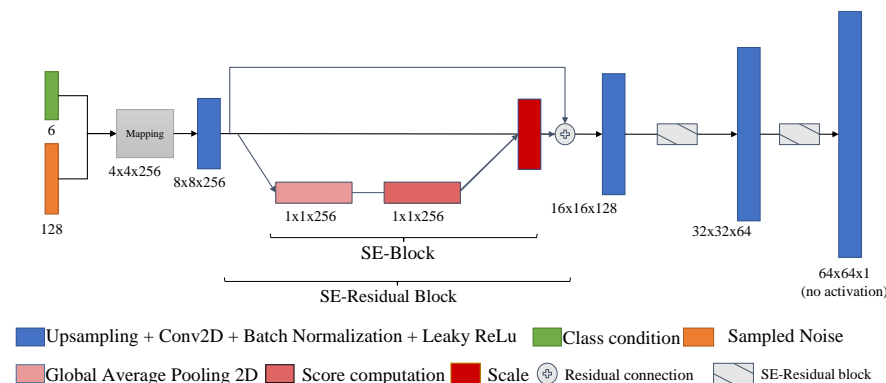


Fig. 1. The schematic illustrates the structure of the proposed SE-ResNet generator for the generation of synthetic mel spectrograms.

The generator has an overall of 1,537,316 parameters. For the discriminator we use a fully convolutional network architecture with a total of 4,321,153

parameters analogous to our own previous work [12]. Both the generator and discriminator employ the LeakyReLU non-linear activation function throughout the whole network structure. As a loss function the Wasserstein Loss with Gradient Penalty (GP) was chosen with GP weight equal to $\lambda = 10$. For both the generator and the discriminator, we utilized the Adam optimizer with a learning rate of $\lambda = 5 \times 10^{-4}$. The discriminator was trained for 5 extra steps per epoch. The implementation and training of all reported results were done using Tensorflow/Keras 2.6 using a Google Cloud instance running a single NVIDIA T4 GPU.

The determination of when to stop the training process is notoriously difficult for the training of GANs. To assess the quality of the generated samples, we repeatedly compute a custom version of the Fréchet Inception Distance [20] which is computed based on the features of the last convolutional layer of a ResNet-18 [19] pre-trained on the THA data set published in [12]. The training process is stopped when the lowest FID is observed which is computed using the equation 2, where μ_r and μ_g is the feature-wise mean of the real and generated spectrograms, C_r and C_g are the covariance matrices.

$$FID = \|\mu_r - \mu_g\|^2 + \text{Tr}(C_r + C_g - 2 * \sqrt{C_r * C_g}) \quad (2)$$

2.3 Classifier for Evaluation

For the evaluation of the proposed improved data augmentation method we employed a ResNet-18 classifier as previously reported in [12] which is a standard convolutional neural network architecture for spectrogram-based audio classification tasks and has been shown to achieve state-of-the-art results in medical acoustic sensing applications [1,12]. To be able to compare the results presented within this work with the previous results, we augment 100% synthetic samples for each class present in the data set. The classifier was trained for 20 epochs using 5-fold cross-validation technique. We used categorical cross-entropy loss with the Adam optimizer and the following hyperparameters: learning rate = 10^{-5} , $\beta_1 = 0.9$, $\beta_2 = 0.99$

3 Results

In order to visually compare the quality of the proposed model, we present per-class randomly selected ground truth data, generated spectrograms from the proposed model, and synthetic spectrograms generated from the previous augmentation framework [12] in figure 2.

We stopped the training by frequently monitoring the quality of the generated samples through the computation of the FID as described in equation 2 and selected the best model with the lowest FID score which was subsequently used to augment the data set by doubling the number of samples for each class, the best performing augmentation strategy identified in previous work. We report the mean Macro F1 score over a five-fold cross validation experiment in the format

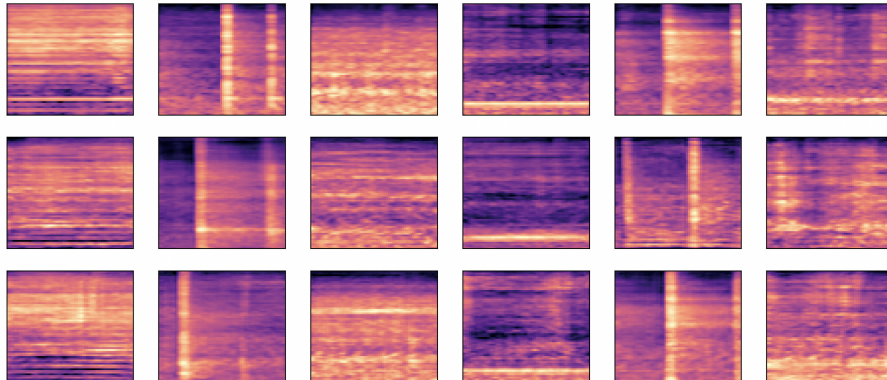


Fig. 2. Log-mel spectrogram of random samples for each class (top row); log-mel spectrogram of random generated images of our proposed model (second row); log-mel spectrogram of the model proposed in the previous work[12] (bottom row). Respective classes from left to right: Sawing, Adjustment, Reaming, Coagulation, Insertion, Suction

mean \pm standard deviation. A comparison between the classifier performance with no augmentations, using classical signal- and spectrogram-processing-based methods, the method proposed in our own previous work [12], and the proposed model is shown in the Table 1.

Augmentation method	FID	Macro F1-Score (mean \pm std)	Relative improvement
No augmentation		93.9 \pm 2.5%	
White noise		92.87 \pm 0.99%	-1.03%
Pitch Shift		94.73 \pm 1.28%	0.83%
Time Stretch		95.0 \pm 1.49%	1.1%
SpecAugment[22]		94.23 \pm 1.14%	0.33%
cWGAN-GP[12]	3.30	95.60 \pm 2.6%	1.7%
Our method	3.01	96.74 \pm 1.03%	2.84%

Table 1. Comparison of different augmentation methods for clinical audio data. All reported results were obtained by applying the respective augmentation method to double the number of samples for each class of the public THA sounds data set.

To analyze the redundancy in learned feature space of the proposed model and compare it with the previously published method, we plot the correlation matrices computed from intermediate layers of the network to analyze the redundancy of features in figure 3. The results show that the redundancy of features is significantly reduced by introducing residual Squeeze and Excitation modules in the generator network.

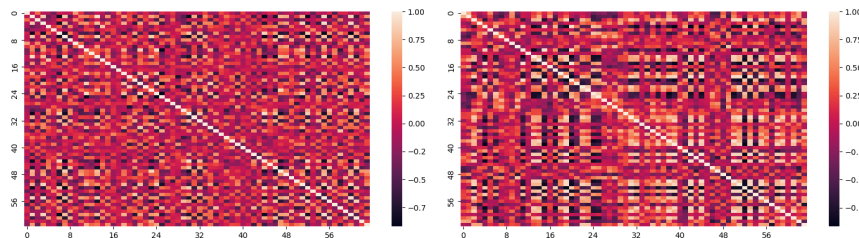


Fig. 3. Sample correlation matrices of features learned by the proposed model (left column) and cWGAN-GP published in previous work [12]. The correlation matrices are computed from different intermediate layers of the generator network. The plots represent the correlation in the feature space after the second-last convolutional layer with dimensions $32 \times 32 \times 64$. The significantly lower correlation values obtained after introducing Squeeze & Excitation block demonstrate the reduced correlation among features and therefore reduced feature redundancy.

4 Discussion

Deep learning-based acoustic sensing has been shown to have high potential for clinical applications in diagnostics and interventional guidance, can be used for multimodal sensing to complement established assistance systems, and provide data beyond the limits of computer aided diagnostic and interventional support systems. However, to achieve state-of-the-art results, learning-based systems rely on big training data sets to generalize well for unseen cases. Obtaining these large amounts of clinical data is a common problem for the design of deep learning-based support and guidance systems in medicine. Advanced augmentation methods have been designed for medical imaging applications [24,25] and a first method for the augmentation of clinical audio data sets has been proposed by the authors in previous work [12].

In the present work, the results show that the proposed method outperforms all previously suggested augmentation methods. In comparison to the first generative modeling based method for clinical audio data, we outperform the model by a margin of 1.14% in Macro F1-Score. While this is an incremental improvement, we could significantly improve the results by only adding a total number of 11232 additional parameters which corresponds to a parameter growth of only 0.74% for the generator model. Furthermore, the correlation analysis of intermediate latent features revealed that the introduced residual Squeeze and Excitation modules reduce the redundancy in the learned features of the generator model. Therefore, the proposed architecture is a highly valuable extension in the generator architecture for an improved synthetic generation of mel spectrograms. An improvement of 0.3 in the reported FID score underlines the capabilities of the proposed architectural modifications.

The proposed approach can generate any arbitrary number of samples for the classes present in the learned data set distribution and could therefore be employed to address data imbalance issues. However, in the current work we focused on improving the quality of the generated samples. Therefore, a more thorough investigation regarding the influence of different augmentation schemes using conditional generative data augmentation should be subject to future research.

By introducing a generative deep-learning method, the processing time for generating the augmentations increases in comparison to simple signal processing-based approaches. To investigate the capabilities of the proposed method, the model should be trained and evaluated on multiple relevant clinical audio data sets in future research.

5 Conclusion

In this work, we propose an enhanced generator architecture for conditional generative learning-based data augmentation of clinical audio data. We outperform all previously published methods and provide an in-depth analysis of the proposed modifications, residual Squeeze and Excitation modules in the generator structure. The method is able to increase the quality of synthetically generated samples by 0.3 in terms of FID score and improves the performance of a classifier trained on the augmented data set by a margin of 2.84% in terms of Macro F1-Score. All presented results are evaluated on a public data set containing sounds of a Total Hip Arthroplasty procedure which was recorded in the real operating room and evaluated using a 5-fold cross validation scheme. The obtained results show that the proposed method has great potential to improve the problem of data limitations for the design of clinical acoustic sensing systems.

Acknowledgement

This work is part of the SURGENT project under the umbrella of Hochschulmedizin Zürich.

References

1. M. Seibold, A. Hoch, D. Suter, M. Farshad, P. O. Zingg, N. Navab, P. Fürnstahl, Acoustic-based spatio-temporal learning for press-fit evaluation of femoral stem implants, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021, pp. 447–456.
2. Q. Goossens, L. Pastrav, J. Roosen, M. Mulier, W. Desmet, J. Vander Sloten, K. Denis, Acoustic analysis to monitor implant seating and early detect fractures in cementless tha: An in vivo study, *Journal of Orthopedic Research* (2020).
3. A. Arami, J.-R. Delaloye, H. Rouhani, B. M. Jolles, K. Aminian, Knee implant loosening detection: A vibration analysis investigation, *Annals of Biomedical Engineering* 46 (2018) 97–107.

4. K. S. Kim, J. H. Seo, J. U. Kang, C. G. Song, An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis, *Computer Methods and Programs in Biomedicine* 94 (2) (2009) 198–206.
5. M. Seibold, S. Maurer, A. Hoch, P. Zingg, M. Farshad, N. Navab, P. Fürnstahl, Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery, *Scientific Reports* 11 (2021).
6. A. Illanes, A. Boese, I. Maldonado, A. Pashazadeh, A. Schaufler, N. Navab, M. Friebe, Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing, *Scientific Reports* 8 (2018).
7. K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith, H. M. Alharthi, W. M. Alghamdi, M. S. Alshahrani, Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities, *IEEE Access* 9 (2021) 102327–102344.
8. N. Giordano, M. Knaflitz, A novel method for measuring the timing of heart sound components through digital phonocardiography, *Sensors* 19 (8) (2019).
9. H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-y. Chang, T. Sainath, Deep learning for audio signal processing, *IEEE Journal on Selected Topics in Signal Processing* 14 (2019) 206–219.
10. V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210.
11. J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound research, in: 22nd ACM International Conference on Multimedia (ACM-MM’14), Orlando, FL, USA, 2014, pp. 1041–1044.
12. M. Seibold, A. Hoch, M. Farshad, N. Navab, P. Fürnstahl, Conditional generative data augmentation for clinical audio datasets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 345–354.
13. J. Liu, B. Zhuang, Z. Zhuang, Y. Guo, J. Huang, J. Zhu, M. Tan, Discrimination-aware network pruning for deep model compression, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
14. I. Joshi, A. Utkarsh, P. Singh, A. Dantcheva, S. D. Roy, P. K. Kalra, On restoration of degraded fingerprints, *Multimedia Tools and Applications* (2022) 1–29.
15. P. Singh, V. K. Verma, P. Rai, V. Namboodiri, Leveraging filter correlations for deep model compression, in: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision, 2020, pp. 835–844.
16. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
17. R. Roy, I. Joshi, A. Das, A. Dantcheva, 3d cnn architectures and attention mechanisms for deepfake detection, in: Handbook of Digital Face Manipulation and Detection, 2022, pp. 213–234.
18. M. Choi, H. Kim, B. Han, N. Xu, K. M. Lee, Channel attention is all you need for video frame interpolation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10663–10671.
19. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
20. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Proceedings

- of the 31st International Conference on Neural Information Processing Systems, 2017, p. 6629–6640.
21. S. S. Stevens, J. Volkman, E. B. Newman, A scale for the measurement of the psychological magnitude pitch, *The Journal of the Acoustical Society of America* 8 (1937).
 22. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, *Interspeech 2019* (2019).
 23. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein gans, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 5769–5779.
 24. M. Tirindelli, C. Eilers, W. Simson, M. Paschali, M. F. Azampour, N. Navab, Rethinking ultrasound augmentation: A physics-inspired approach, in: *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 690–700.
 25. H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in: A. Gooya, O. Goksel, I. Oguz, N. Burgos (Eds.), *Simulation and Synthesis in Medical Imaging*, 2018, pp. 1–11.

Part III

Conclusions and Outlook

Summary of Findings

The work presented in this dissertation underlines the potential of acoustic sensing for medical applications and proposes novel deep learning-based solutions for unmet clinical problems in orthopedic surgery, i.e. a system for error prevention during hand-held surgical drilling, a method for the assessment of the optimal insertion endpoint during femoral stem insertion, and a new approach for the detection of pedicle screw loosening using active excitation vibration analysis. To address the issue of limited data for the proposed learning-based methods, we introduce a novel generative model-based data augmentation method that is able to generate realistic class-conditioned samples from a learned data set distribution and improve the performance of a deep learning-based acoustic sensing system. The presented work shows that acoustic signals have great potential for the development of novel multimodal sensing solutions for medical applications and can provide solutions for problems where conventional systems, such as surgical navigation or medical imaging reach their limits. The following paragraphs summarize the main findings of this dissertation.

Clinical Applications | Through a thorough literature review and the evaluation of the systems proposed within in the scope of this thesis, orthopedics has been identified as a promising field for the deployment of acoustic sensing systems, as the interaction of physicians with the musculoskeletal system and surgical tools often creates distinctive audible noise. In discussions with experienced orthopedic surgeons during the course of the author's PhD studies, the physicians reported that they implicitly use the distinctive sounds generated e.g. by tool-tissue interaction to infer additional information about interventions, such as the quality of screw hold during screw insertion, the quality of bone during hammering, or the seating of an implant during insertion. These measurements cannot be obtained only by conventional computer aided surgery systems but the information is contained in air-borne and structure-borne acoustic signals which can be exploited for the design of medical acoustic sensing systems. Acoustic sensing solutions have, apart from orthopedic use cases, also been utilized in first applications outside of the orthopedic field and show great potentials for other applications, e.g. for soft tissue differentiation in minimally invasive surgery as presented in the author's own previous work [78], or needle guidance in soft tissues [38]. Furthermore, there is still great potential to explore other applications in orthopedics where distinctive sounds generated through tool-tissue interaction or active excitation of bony structures can be exploited to design automated acoustic sensing solutions for improved patient care.

Data Requirements | To acquire training data for deep learning-based medical acoustic sensing systems, it is crucial to access a realistic environment and test the system under relevant conditions. Therefore, for the evaluation of the methods proposed within this dissertation, the proposed systems have been evaluated in human cadaveric setups (the works presented in sections 3.2.1, 3.3.1, and 3.2.2) or tested with real acoustic signals captured within the operating room at Balgrist University Hospital (the work presented in section 3.5.1). However, the capturing of data in a realistic medical environment such as the

operating room requires special considerations to ensure the unhindered execution of the medical procedure and imposes the need for ethical and regulatory approvals to fulfill all regulations regarding patient consent and usage of personalized patient data. Furthermore, human cadaver experiments are associated with extensive preparation, access to specialized infrastructure, and high financial costs. Therefore, compromises have to be made to evaluate novel algorithms and solutions on limited data sets. To address this issue, we propose a novel data augmentation method to artificially increase the size of a limited data set by using a generative adversarial network to sample new data points from a learned data distribution and show that our method improves the results, as well as stabilizes the training. Furthermore, signals of interest often occur scarcely in medical applications, e.g. drill breakthrough events in relation to regular drilling events which results in the acquisition of imbalanced data sets. To address this issue, the paper presented in section 3.2.1 employs the focal loss function during model training to adapt the proposed method to imbalanced data. In this work, we show that this approach can drastically improve the performance of the classifier in the presence of class imbalance.

Signal Acquisition | The acquisition of high-quality, low-noise, and relevant acoustic signals is essential to design acoustic sensing solutions for medical applications. To capture structure-borne acoustic signals from the area of operation, we developed a custom and modular contact microphone setup which can be attached to tools or directly to the patient's or specimens' skin surface. Because of its modular structure consisting of a replaceable low-cost microphone connected via rigid connectors to a pre-amplification stage, the system could easily be integrated into clinical workflows, but currently does not have medical certification. Furthermore, the piezo-based contact microphones used in the papers presented in sections 3.3.1, 3.2.2, and 3.2.1 have the advantage of only capturing structure-borne vibrations from the object that they are attached to, they do not pick up any environment noise. The system is described in detail in the publication presented in section 3.2.1. Additionally, we developed a safe setup using a rigid stand with counterweights and a directed condenser microphone to capture a data set from real THA procedures which is described in publication 3.5.1. The setup enables a high-quality audio capture from the operating field and was approved for data acquisition in the operating room by an ethics committee.

Feature Representations | All presented solutions are designed on the basis of mel spectrograms which have been identified in previous work as the feature of choice for deep learning-based acoustic signal processing tasks. Especially in low-data regimes, spectrograms provide a compact representation while capturing the relevant information and signal characteristics for further analysis. Furthermore, spectrograms are well suited as input feature for established feature extraction models like convolutional neural networks. This combination is a solid basis for the development of customized machine learning solutions and network architectures to address the individual challenges of specific clinical problems.

Open Source | To encourage further research in the direction of acoustic sensing for medical applications, the availability of code and data is very helpful for other researchers to develop novel methods based on previous work and benchmark their proposed systems in comparative studies with previously published work. Therefore, code repositories, data sets and pretrained

models for the works presented within this dissertation have partially been made publicly available ¹.

Summary | In summary, it can be stated that acoustic sensing has been shown to have great potential for various medical applications in diagnostic and interventional contexts, for surgical guidance, surgical error prevention, intraoperative decision support, and to develop novel sensing paradigms for medical diagnostics. Acoustic signals can be used complementary to surgical navigation and optical tracking systems, can be employed where conventional computer aided surgery systems reach their limits, and provide additional multimodal information about the surgical scene which can be beneficial for open, minimally invasive, and robotic surgery. They can also be exploited to obtain additional information in diagnostic contexts, e.g. about the condition of orthopedic implants, complementary to established methods such as medical imaging.

¹<https://rocs.balgrist.ch/open-access/>

Future Directions

Based on the foundation laid by this work, we hope that we can encourage the research community to further investigate the usage of acoustic signals in the medical domain, extend and advance the algorithms for audio signal processing, and identify novel and exciting applications for sound and vibration analysis in medical diagnostic and interventional contexts. The most promising directions for future research which have been identified within the course of this dissertation are discussed in the following paragraphs.

Clinical Translation | Most systems proposed in the related literature were developed and evaluated in experimental setups and controlled lab environments. Some of them, including the methods developed within the scope of this thesis, have been evaluated in human cadaveric experimental setups which provides a more realistic environment compared to artificial bone models or animal specimens. Real data from the operating room or patients was used only in few projects, e.g. the analysis of joint noise from patients for vibroarthrography as described in section 1.3.1 or the analysis of air-borne sounds from the operating room during THA surgery by Goossens et al. [32]. While being a simplified environment, different abstractions of reality in an experimental setup can be utilized to show the feasibility of a system and have the advantage of less regulatory requirements and preparation efforts. However, for the translation of acoustic sensing solutions into clinical use, many additional considerations have to be taken into account. A major point is the sterilization of the contact microphones which is crucial when the devices are in contact with the patient anatomy and surgical tools used in the operating field. To solve this challenge, the design of the contact microphones has to be improved and adapted towards sterilizability, where the modular design of the signal acquisition hardware would be beneficial for the integration of a sterilizable device into the current acquisition setup. As the contact microphones used within our experiments are not a medical product, additional tests, e.g. electromagnetic interference and risk assessment, have to be performed to ensure patient safety in every possible situation. Finally, a medical certification has to be obtained to use the device with real patients in the clinic. For air-borne microphones, the regulatory considerations are less strict to capture acoustic signals in a real environment, as they are not in contact with the patient or tools. On the other hand, they capture additional noise from the room, e.g. noise from other medical devices and staff conversations which imposes the need to take personalized data protection into account.

Technical Directions | Because deep learning research is advancing at an incredible pace, future work should constantly integrate and evaluate novel techniques from basic machine learning research for the design of novel medical acoustic sensing systems. The following section will give an overview about the most promising short-term directions.

Convolutional neural networks have been extensively used for feature extraction from grid-like domains like images and also spectrograms for almost a decade and various enhancements, such as residual connections or Squeeze & Excitation modules, have been proposed in the

literature which have been shown to improve the performance also for spectrogram-based audio classification tasks in the works presented within the scope of this thesis. Despite the fact that these enhanced CNN-based models are powerful feature extractors, novel attention-based feature extraction techniques like the Vision Transformer (ViT) have recently gained popularity and have been applied to a variety of problems. However, even though the training of transformer-based models from scratch usually require large amounts of training data or large-scale pretraining, first results with optimized ViT-based models have been published that outperform similar size ResNets [15] without pretraining or extensive data augmentations. In recently published work, first applications of transformer-based architectures have also achieved promising results for spectrogram-based audio processing tasks [30, 124]. Also for sequence-based models like the system proposed in 3.3.1, attention-based sequence modules have to be investigated for potential performance improvements.

Data Requirements | A main challenge for the application of deep learning-based acoustic sensing systems in the medical domain is the acquisition of sufficient amounts of relevant and high-quality data. Because data capture in realistic environments such as human cadaveric experiments or in the real operating room is associated with extensive effort, cost, and regulatory matters, data augmentation can be an important tool to artificially enlarge the size of a training data set to improve the results and generalization of a deep learning-based system. Even though first promising results have been obtained within the scope of the present thesis (as described in the publications in sections 3.5.1 and 3.5.2), the proposed data augmentation method should be further improved and extended, and a thorough comparison with other augmentation methods and their combination for an optimized data augmentation scheme should be performed.

All models proposed within the works of the present thesis have been trained from scratch, as large-scale audio data sets and pretrained feature extractors for spectrogram features are scarce in comparison to pretrained models for computer vision applications, e.g. models pretrained on ImageNet [92] which can be utilized for transfer learning-based downstream tasks with frozen weights or fine-tuning. While there are a few larger audio data sets, like YouTube-8M [2] or Urbansound-8k [95] publicly available, the medical domain lacks public large-scale audio data sets which could be utilized by the community to develop new solutions based on transfer learning. Within the scope of the thesis, first medical audio data sets have been published or shared with the research community and have already been used by other researchers for the evaluation of their proposed system [113]. Therefore, future research should encourage the creation, publication, and curation of large-scale, multi-center data sets to develop better algorithms and achieve improved generalization through large-scale data and increased diversity in the data set distribution of medical audio data sets.

Even though methods for the handling of imbalanced data have been identified in the context of the present thesis, such as using alternative loss functions to specifically address the problem of class imbalance, future research should address these challenges and provide novel, improved solutions for limited and imbalanced data.

Outlook | Even though there is already a body of related work for the usage of automated acoustic sensing systems in a wide variety of medical applications, we are only at the beginning of using structure- and air-borne acoustic signals for multimodal sensing in medical diagnostics and interventions. There is great potential not only in terms of surgical error prevention, surgical decision support in computer aided surgery, and multimodal sensing

for diagnostics. An application where acoustic signals can play an important role in future is workflow recognition and surgical phase detection where camera views can be blocked by staff members in open surgery or impaired by smoke in minimally invasive surgery or water in arthroscopy. Here, acoustic signals can provide complementary information for the development of multimodal phase recognition systems that can be utilized for a variety of downstream tasks such as surgical education or skill assessment. We present preliminary results for the usage of acoustic signals in surgical workflow recognition in section 3.5.1 which motivates further research in this direction.

Furthermore, we identified smart orthopedic implants as a promising direction for further research, where an implant could sense a potential implant loosening itself based on the analysis of structure-borne vibrations with a miniature sensor system integrated into the implant structure. However, to realize this sensor enhancement, additional challenges such as the development of a miniaturized and integrated sensor system and the problems of energy supply and signal transmission have to be solved in future research.

As robotic surgery becomes increasingly important to support surgeons in performing highly delicate surgical actions and robotic surgery systems will also (partly) operate autonomously in future, multimodal sensing approaches are an essential building block for robot perception, where air- and structure-borne acoustic signals can play an important role. The European research project FAROS¹, which is conducted within our group at Balgrist University Hospital together with partner institutions, follows up on this idea and has the goal of improving the functional accuracy of robotic surgery through the embedding of multimodal sensing and physical intelligence into surgical robotic systems.

Finally, we believe that acoustic sensing should be considered as an integral part of not only future robotic systems, but also in general for intelligent systems, e.g. in Augmented Reality systems such as Head Mounted Devices (HMDs), to enable these systems to create a better internal digital representation of the real world through multimodal sensing. The human hearing is one of our most important senses and connects us to the physical world by informing us about the situation around us and extending the information we perceive through visual cues. Sounds communicate with the human brain much quicker than visual information, are essential for communication, and can alert us about events that happen outside of our field of view, even during sleep. This analogy underlines the potential of acoustic sensing to complement established computer vision systems and other sensing modalities as an integral part for intelligent systems of the future.

¹<https://h2020faros.eu/>, European Union's Horizon 2020, Grant No. 101016985

Part IV

Appendix

Abstracts of Publications not Discussed in this Thesis

Towards a Low-Cost Monitor-Based Augmented Reality Training Platform for At-Home Ultrasound Skill Development

Marine Y. Shao, Tamara Vagg, **Matthias Seibold**, Mitchell Doughty

Ultrasound education traditionally involves theoretical and practical training on patients or on simulators; however, difficulty accessing training equipment during the COVID-19 pandemic has highlighted the need for home-based training systems. Due to the prohibitive cost of ultrasound probes, few medical students have access to the equipment required for at home training. Our proof of concept study focused on the development and assessment of the technical feasibility and training performance of an at-home training solution to teach the basics of interpreting and generating ultrasound data. The training solution relies on monitor-based augmented reality for displaying virtual content and requires only a marker printed on paper and a computer with webcam. With input webcam video, we performed body pose estimation to track the student's limbs and used surface tracking of printed fiducials to track the position of a simulated ultrasound probe. The novelty of our work is in its combination of printed markers with marker-free body pose tracking. In a small user study, four ultrasound lecturers evaluated the training quality with a questionnaire and indicated the potential of our system. The strength of our method is that it allows students to learn the manipulation of an ultrasound probe through the simulated probe combined with the tracking system and to learn how to read ultrasounds in B-mode and Doppler mode.

Journal of Imaging, 8(11), 305, 2022.

HAPPY: Hip Arthroscopy Portal Placement Using Augmented Reality

Tianyu Song, Michael Sommersperger, The Anh Baran, **Matthias Seibold**, and Nassir Navab

Correct positioning of the endoscope is crucial for successful hip arthroscopy. Only with adequate alignment can the anatomical target area be visualized and the procedure be successfully performed. Conventionally, surgeons rely on anatomical landmarks such as bone structure, and on intraoperative X-ray imaging, to correctly place the surgical trocar and insert the endoscope to gain access to the surgical site. One factor complicating the placement

is deformable soft tissue, as it can obscure important anatomical landmarks. In addition, the commonly used endoscopes with an angled camera complicate hand–eye coordination and, thus, navigation to the target area. Adjusting for an incorrectly positioned endoscope prolongs surgery time, requires a further incision and increases the radiation exposure as well as the risk of infection. In this work, we propose an augmented reality system to support endoscope placement during arthroscopy. Our method comprises the augmentation of a tracked endoscope with a virtual augmented frustum to indicate the reachable working volume. This is further combined with an in situ visualization of the patient anatomy to improve perception of the target area. For this purpose, we highlight the anatomy that is visible in the endoscopic camera frustum and use an automatic colorization method to improve spatial perception. Our system was implemented and visualized on a head-mounted display. The results of our user study indicate the benefit of the proposed system compared to baseline positioning without additional support, such as an increased alignment speed, improved positioning error and reduced mental effort. The proposed approach might aid in the positioning of an angled endoscope, and may result in better access to the surgical area, reduced surgery time, less patient trauma, and less X-ray exposure during surgery.

Journal of Imaging, 8(11), 302, 2022.

Sonification as a Reliable Alternative to Conventional Visual Surgical Navigation

Sasan Matinfar, Mehrdad Salehi, Daniel Suter, **Matthias Seibold**, Navid Navab, Shervin Dehghani, Florian Wanivenhaus, Philipp Fürnstahl, Mazda Farshad, Nassir Navab

Despite the undeniable advantages of image-guided surgical assistance systems in terms of accuracy, such systems have not yet fully met surgeons' needs or expectations regarding usability, time efficiency, and their integration into the surgical workflow. On the other hand, perceptual studies have shown that presenting independent but causally correlated information via multimodal feedback involving different sensory modalities can improve task performance. This article investigates an alternative method for computer-assisted surgical navigation, introduces a novel sonification methodology for navigated pedicle screw placement, and discusses advanced solutions based on multisensory feedback. The proposed method comprises a novel sonification solution for alignment tasks in four degrees of freedom based on frequency modulation (FM) synthesis. We compared the resulting accuracy and execution time of the proposed sonification method with visual navigation, which is currently considered the state of the art. We conducted a phantom study in which 17 surgeons executed the pedicle screw placement task in the lumbar spine, guided by either the proposed sonification-based or the traditional visual navigation method. The results demonstrated that the proposed method is as accurate as the state of the art while decreasing the surgeon's need to focus on visual navigation displays instead of the natural focus on surgical tools and targeted anatomy during task execution.

arXiv preprint arXiv:2206.15291, 2022

Towards markerless surgical tool and hand pose estimation

Jonas Hein*, **Matthias Seibold***, Federica Bogo, Marc Pollefeys, Mazda Farshad, Philipp Frnstahl, Nassir Navab (* equal contribution)

Purpose: Tracking of tools and surgical activity is becoming more and more important in the context of computer assisted surgery. In this work, we present a data generation framework, dataset and baseline methods to facilitate further research in the direction of markerless hand and instrument pose estimation in realistic surgical scenarios.

Methods: We developed a rendering pipeline to create inexpensive and realistic synthetic data for model pretraining. Subsequently, we propose a pipeline to capture and label real data with hand and object pose ground truth in an experimental setup to gather high-quality real data. We furthermore present three state-of-the-art RGB-based pose estimation baselines.

Results: We evaluate three baseline models on the proposed datasets. The best performing baseline achieves an average tool 3D vertex error of 16.7 mm on synthetic data as well as 13.8 mm on real data which is comparable to the state-of-the-art in RGB-based hand/object pose estimation.

Conclusion: To the best of our knowledge, we propose the first synthetic and real data generation pipelines to generate hand and object pose labels for open surgery. We present three baseline models for RGB based object and object/hand pose estimation based on RGB frames. Our realistic synthetic data generation pipeline may contribute to overcome the data bottleneck in the surgical domain and can easily be transferred to other medical applications.

International Journal of Computer Assisted Radiology and Surgery, 16, pp. 799–808, 2021

Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions

Daniel Ostler*, **Matthias Seibold***, Jonas Fuchtmann, Nicole Samm, Hubertus Feussner, Dirk Wilhelm, Nassir Navab (* equal contribution)

Purpose: Minimally invasive surgery (MIS) has become the standard for many surgical procedures as it minimizes trauma, reduces infection rates and shortens hospitalization. However, the manipulation of objects in the surgical workspace can be difficult due to the unintuitive handling of instruments and limited range of motion. Apart from the advantages of robot-assisted systems such as augmented view or improved dexterity, both robotic and MIS techniques introduce drawbacks such as limited haptic perception and their major reliance on visual perception.

Methods: In order to address the above-mentioned limitations, a perception study was conducted to investigate whether the transmission of intra-abdominal acoustic signals can potentially improve the perception during MIS. To investigate whether these acoustic signals

can be used as a basis for further automated analysis, a large audio data set capturing the application of electrosurgery on different types of porcine tissue was acquired. A sliding window technique was applied to compute log-mel-spectrograms, which were fed to a pre-trained convolutional neural network for feature extraction. A fully connected layer was trained on the intermediate feature representation to classify instrument–tissue interaction.

Results: The perception study revealed that acoustic feedback has potential to improve the perception during MIS and to serve as a basis for further automated analysis. The proposed classification pipeline yielded excellent performance for four types of instrument–tissue interaction (muscle, fascia, liver and fatty tissue) and achieved top-1 accuracies of up to 89.9%. Moreover, our model is able to distinguish electrosurgical operation modes with an overall classification accuracy of 86.40

Conclusion: Our proof-of-principle indicates great application potential for guidance systems in MIS, such as controlled tissue resection. Supported by a pilot perception study with surgeons, we believe that utilizing audio signals as an additional information channel has great potential to improve the surgical performance and to partly compensate the loss of haptic feedback.

International Journal of Computer Assisted Radiology and Surgery, 15, pp. 771–779, 2020

pix2xray: converting RGB images into X-rays using generative adversarial networks

Mustafa Haiderbhai, Sergio Ledesma, Sing Chun Lee, **Matthias Seibold**, Phillipp Frnstahl, Nassir Navab, Pascal Fallavollita

Purpose: We propose a novel methodology for generating synthetic X-rays from 2D RGB images. This method creates accurate simulations for use in non-diagnostic visualization problems where the only input comes from a generic camera. Traditional methods are restricted to using simulation algorithms on 3D computer models. To solve this problem, we propose a method of synthetic X-ray generation using conditional generative adversarial networks (CGANs).

Methods: We create a custom synthetic X-ray dataset generator to generate image triplets for X-ray images, pose images, and RGB images of natural hand poses sampled from the NYU hand pose dataset. This dataset is used to train two general-purpose CGAN networks, pix2pix and CycleGAN, as well as our novel architecture called pix2xray which expands upon the pix2pix architecture to include the hand pose into the network.

Results: Our results demonstrate that our pix2xray architecture outperforms both pix2pix and CycleGAN in producing higher-quality X-ray images. We measure higher similarity metrics in our approach, with pix2pix coming in second, and CycleGAN producing the worst results. Our network performs better in the difficult cases which involve high occlusion due to occluded poses or large rotations.

Conclusion: Overall our work establishes a baseline that synthetic X-rays can be simulated using 2D RGB input. We establish the need for additional data such as the hand pose to produce clearer results and show that future research must focus on more specialized architectures to improve overall image clarity and structure.

International Journal of Computer Assisted Radiology and Surgery, 15, pp. 973-980, 2020

Pivot calibration concept for sensor attached mobile c-arms

Sing Chun Lee*, **Matthias Seibold***, Philipp Fürnstahl, Mazda Farshad, Nassir Navab (* equal contribution)

Medical augmented reality has been actively studied for decades and many methods have been proposed to revolutionize clinical procedures. One example is the camera augmented mobile C-arm (CAMC), which provides a real-time video augmentation onto medical images by rigidly mounting and calibrating a camera to the imaging device. Since then, several CAMC variations have been suggested by calibrating 2D/3D cameras, trackers, and more recently a Microsoft HoloLens to the C-arm. Different calibration methods have been applied to establish the correspondence between the rigidly attached sensor and the imaging device. A crucial step for these methods is the acquisition of X-Ray images or 3D reconstruction volumes; therefore, requiring the emission of ionizing radiation. In this work, we analyze the mechanical motion of the device and propose an alternative method to calibrate sensors to the C-arm without emitting any radiation. Given a sensor is rigidly attached to the device, we introduce an extended pivot calibration concept to compute the fixed translation from the sensor to the C-arm rotation center. The fixed relationship between the sensor and rotation center can be formulated as a pivot calibration problem with the pivot point moving on a locus. Our method exploits the rigid C-arm motion describing a Torus surface to solve this calibration problem. We explain the geometry of the C-arm motion and its relation to the attached sensor, propose a calibration algorithm and show its robustness against noise, as well as trajectory and observed pose density by computer simulations. We discuss this geometric-based formulation and its potential extensions to different C-arm applications.

SPIE Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling 2020, Houston, TX, USA, 2020

Sonification for Process Monitoring in Highly Sensitive Surgical Tasks

Sasan Matinfar, Thomas Hermann, Matthias Seibold, Philipp Fürnstahl, Mazda Farshad, Nassir Navab

Surgeons usually have to keep track of many variables during a surgical intervention. This paper introduces three novel sonification approaches for fluid-related process data monitoring in the highly sensitive surgical context. From the instantaneous fluid (creation or expenditure)

rate, a number of feature time series has been computed, including the cumulative fluid volume or filtered signals, which are in turn used for a set of sonification methods that structure a composed soundscape, either natural, musical or hybrid in real-time. We present 3 variations of 3 approaches with introductory example videos, each followed by results of a first user study in search of the preferred/most acceptable auditory representations. The qualitative evaluation of our method shows the potentials for further research in this field.

Nordic Sound and Music Computing Conference 2019 (Nordic SMC 2019), Stockholm, Sweden, 2019

Bibliography

- [1] J. Abeßer. “A Review of Deep Learning Based Methods for Acoustic Scene Classification”. In: *Applied Sciences* 10.6 (2020) (cit. on p. 12).
- [2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. “YouTube-8M: A Large-Scale Video Classification Benchmark”. In: *CoRR* abs/1609.08675 (2016). arXiv: 1609.08675 (cit. on pp. 63, 94).
- [3] J. Ackermann, M. Wieland, A. Hoch, R. Ganz, J. G. Snedeker, M. R. Oswald, M. Pollefeys, P. O. Zingg, H. Esfandiari, and P. Fürnstahl. “A New Approach to Orthopedic Surgery Planning Using Deep Reinforcement Learning and Simulation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. 2021, pp. 540–549 (cit. on p. 12).
- [4] H. Albin Lomami, C. Damour, G. Rosi, A.-S. Poudrel, A. Dubory, C.-H. Flouzat-Lachaniette, and G. Haiat. “Ex vivo estimation of cementless femoral stem stability using an instrumented hammer”. In: *Clinical Biomechanics* 76 (2020), p. 105006 (cit. on p. 7).
- [5] Z. Alperovich, G. Yamin, E. Elul, G. Bialolenker, and A. A. Ishaaya. “In situ tissue classification during laser ablation using acoustic signals”. In: *Journal of Biophotonics* 12 (9 2019) (cit. on p. 11).
- [6] A. Arami, J.-R. Delaloye, H. Rouhani, B. M. Jolles, and K. Aminian. “Knee Implant Loosening Detection: A Vibration Analysis Investigation”. In: *Annals of Biomedical Engineering* 46 (2018), pp. 97–107 (cit. on p. 9).
- [7] N. Befruoi, J. Elsner, A. Flessner, J. Huvanandana, O. Jarrousse, T. N. Le, M. Müller, W. H. W. Schulze, S. Taing, and S. Weidert. “Vibroarthrography for early detection of knee osteoarthritis using normalized frequency features”. In: *Medical & Biological Engineering & Computing* 56 (2018), pp. 1499–1514 (cit. on pp. 9, 25).
- [8] Y. Bi, M. Lv, C. Song, W. Xu, N. Guan, and W. Yi. “AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life”. In: *IEEE Sensors Journal* 16.3 (2016), pp. 806–816 (cit. on p. 9).
- [9] I. Boesnach, M. Hahn, J. Moldenauer, and T. Beth. “Analysis of Drill Sound in Spine Surgery”. In: *Perspective in Image-Guided Surgery* (2004), pp. 77–84 (cit. on p. 10).
- [10] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric Deep Learning: Going beyond Euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42 (cit. on p. 11).
- [11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901 (cit. on p. 12).

- [12] S. Chandrakala and S. L. Jayalakshmi. “Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance: A Survey and Comparative Studies”. In: *ACM Comput. Surv.* 52.3 (2019) (cit. on p. 12).
- [13] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan. “Data Augmentation Using GANs for Speech Emotion Recognition”. In: *Proc. Interspeech 2019*. 2019, pp. 171–175 (cit. on p. 63).
- [14] C. Chen, T. Sühn, M. Kalmar, I. Maldonado, C. Wex, R. Croner, A. Boese, M. Friebe, and A. Illanes. “Texture differentiation using audio signal analysis with robotic interventional instruments”. In: *Computers in Biology and Medicine* 112 (2019) (cit. on p. 11).
- [15] X. Chen, C.-J. Hsieh, and B. Gong. “When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations”. In: *International Conference on Learning Representations - ICLR 2022*. 2022 (cit. on p. 94).
- [16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”. In: *NIPS 2014 Deep Learning and Representation Learning Workshop*. 2014 (cit. on p. 11).
- [17] J. W. Cooley and J. W. Tukey. “An Algorithm for the Machine Calculation of Complex Fourier Series”. In: *Mathematics of Computation* 19 (90 1965), pp. 297–301 (cit. on p. 20).
- [18] Y. Dai, Y. Xue, and J. Zhang. “State identification based on sound analysis during surgical milling process”. In: *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2015 (cit. on p. 10).
- [19] Y. Dai, Y. Xue, and J. Zhang. “Tissue discrimination based on vibratory sense in robot-assisted spine surgery”. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 2015, pp. 4717–4722 (cit. on p. 7).
- [20] Y. Dai, X. Yuan, and J. Zhang. “Milling State Identification Based on Vibration Sense of a Robotic Surgical System”. In: *IEEE Transactions on Industrial Electronics* 63 (10 2016) (cit. on p. 10, 25).
- [21] S. Davis and P. Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 28 (1980), pp. 357–366 (cit. on p. 21).
- [22] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam. “Classification of audio signals using SVM and RBFNN”. In: *Expert Systems with Applications* 36.3, Part 2 (2009), pp. 6069–6075 (cit. on p. 25).
- [23] A. Dubory, G. Rosi, A. Tijou, H. A. Lomami, C.-H. Flouzat-Lachaniette, and G. Haiat. “A cadaveric validation of a method based on impact analysis to monitor the femoral stem insertion”. In: *Journal of the Mechanical Behavior of Biomedical Materials* 103 (2020) (cit. on p. 10).
- [24] D. Dutta, P. Agrawal, and S. Ganapathy. “A Multi-Head Relevance Weighting Framework for Learning Raw Waveform Audio Representations”. In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2021, pp. 191–195 (cit. on p. 63).
- [25] H. Ewald, U. Timm, C. Ruther, W. Mittelmeier, R. Bader, and D. Kluess. “Acoustic sensor system for loosening detection of hip implants”. In: *2011 Fifth International Conference on Sensing Technology*. 2011, pp. 494–497 (cit. on p. 8).
- [26] H. Fabelo, M. Halicek, S. Ortega, A. Szolna, J. Morera, R. Sarmiento, G. M. Callico, and B. Fei. “Surgical aid visualization system for glioblastoma tumor identification based on deep learning and in-vivo hyperspectral images of human patients”. In: *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by B. Fei and C. A. Linte. Vol. 10951. International Society for Optics and Photonics. SPIE, 2019, p. 1095110 (cit. on p. 7).
- [27] A. J. Fitzpatrick, G. W. Rodgers, G. J. Hooper, and T. B. Woodfield. “Development and validation of an acoustic emission device to measure wear in total hip replacements in-vitro and in-vivo”. In: *Biomedical Signal Processing and Control* 33 (2017), pp. 281–288 (cit. on p. 8).

- [28] C. B. Frank, R. M. Rangayyan, and G. D. Bell. “Analysis of knee joint sound signals for non-invasive diagnosis of cartilage pathology”. In: *IEEE Engineering in Medicine and Biology Magazine* 9 (1 1990), pp. 65–68 (cit. on p. 9).
- [29] H. Ghayvat, S. Pandya, and A. Patel. “Deep Learning Model for Acoustics Signal Based Preventive Healthcare Monitoring and Activity of Daily Living”. In: *2nd International Conference on Data, Engineering and Applications (IDEA)*. 2020, pp. 1–7 (cit. on p. 9).
- [30] Y. Gong, Y.-A. Chung, and J. Glass. “AST: Audio Spectrogram Transformer”. In: *INTERSPEECH 2021*. 2021 (cit. on p. 94).
- [31] Q. Goossens, S. Leuridan, P. Henyš, J. Roosen, L. Pastrav, M. Mulier, W. Desmet, K. Denis, and J. Vander Sloten. “Development of an acoustic measurement protocol to monitor acetabular implant fixation in cementless total hip Arthroplasty: A preliminary study”. In: *Medical Engineering & Physics* 49 (2017), pp. 28–38 (cit. on p. 10).
- [32] Q. Goossens, L. Pastrav, J. Roosen, M. Mulier, W. Desmet, J. Vander Sloten, and K. Denis. “Acoustic analysis to monitor implant seating and early detect fractures in cementless THA: An in vivo study”. In: *Journal of Orthopedic Research* (2020) (cit. on pp. 10, 93).
- [33] D. Griffin and J. Lim. “Signal estimation from modified short-time Fourier transform”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), pp. 236–243 (cit. on p. 22).
- [34] F. Guan, Y. Sun, X. Qi, Y. Hu, G. Yu, and J. Zhang. “State Recognition of Bone Drilling Based on Acoustic Emission in Pedicle Screw Operation”. In: *Sensors (Basel)* 18 (5 2019) (cit. on p. 10).
- [35] P. Gupta, M. J. Moghimi, Y. Jeong, D. Gupta, O. T. Inan, and F. Ayazi. “Precision wearable accelerometer contact microphones for longitudinal monitoring of mechano-acoustic cardiopulmonary signals”. In: *npj Digital Medicine* 3 (2020) (cit. on p. 9).
- [36] J. Hein, M. Seibold, F. Bogo, M. Farshad, M. Pollefeys, P. Fürnstahl, and N. Navab. “Towards markerless surgical tool and hand pose estimation”. In: *International Journal of Computer Assisted Radiology and Surgery* 16 (2021), pp. 799–808 (cit. on p. 2).
- [37] S. Hochreiter and J. Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 11).
- [38] A. Illanes, A. Boese, I. Maldonado, A. Pashazadeh, A. Schaufler, N. Navab, and M. Friebe. “Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing”. In: *Scientific Reports* 8 (2018) (cit. on pp. 10, 25, 89).
- [39] A. Illanes, T. Sühn, N. Esmaili, I. Maldonado, A. Schaufler, C.-H. Chen, A. Boese, and M. Friebe. “Surgical Audio Guidance SurAG: Extracting Non-Invasively Meaningful Guidance Information During Minimally Invasive Procedures”. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. 2019, pp. 567–570 (cit. on p. 11).
- [40] S. Ji, J. Luo, and X. Yang. “A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions”. In: *arXiv e-prints* (2020) (cit. on p. 13).
- [41] L. Jiqing, D. Yuan, H. Jun, Z. Xianyu, and W. Haila. “Sports audio classification based on MFCC and GMM”. In: *2009 2nd IEEE International Conference on Broadband Network & Multimedia Technology*. 2009, pp. 482–485 (cit. on p. 25).
- [42] L. Joskowicz and E. J. Hazan. “Computer-Aided Orthopedic Surgery: Incremental Shift or Paradigm Change?” In: *Medical Image Analysis* 33 (2016), pp. 84–90 (cit. on p. 7).
- [43] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596 (2021), pp. 583–589 (cit. on p. 12).

- [44] L. Khokhlova, D.-S. Komaris, S. Tedesco, and B. O'Flynn. "Assessment of Hip and Knee Joints and Implants Using Acoustic Emission Monitoring: A Scoping Review". In: *IEEE Sensors* 21 (13 2021), pp. 14379–14388 (cit. on p. 8).
- [45] K. S. Kim, J. H. Seo, J. U. Kang, and C. G. Song. "An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis". In: *Computer Methods and Programs in Biomedicine* 94.2 (2009), pp. 198–206 (cit. on p. 9).
- [46] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*. 2021 (cit. on p. 12).
- [47] S. Krishnan, R. M. Rangayyan, G. D. Bell, C. B. Frank, and K. O. Ladly. "Adaptive filtering, modelling and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology". In: *Medical and Biological Engineering and Computing* 35 (1997), pp. 677–684 (cit. on p. 9).
- [48] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc., 2012 (cit. on p. 11).
- [49] E. A. Krupinski. "Current perspectives in medical image perception". In: *Attention, Perception, & Psychophysics* 72 (5 2010) (cit. on p. 7).
- [50] R. T. H. Laennec. "De L'Auscultation Médiante; ou Traité du Diagnostic des Maladies des Poumons et du Cœur, fondé principalement sur ce Nouveau Moyen d'Exploration". In: *Edinburgh Medical and Surgical Journal* 18 (72 1822), pp. 447–474 (cit. on p. 5).
- [51] J. Laguarda, F. Hueto, and B. Subirana. "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings". In: *IEEE Open Journal of Engineering in Medicine and Biology* 1 (2020), pp. 275–281 (cit. on pp. 8, 13).
- [52] J. Lai, R. Woodward, Y. Alexandrov, Q. ain Munnee, C. C. Lees, R. Vaidyanathan, and N. C. Nowlan. "Performance of a wearable acoustic system for fetal movement discrimination". In: *PLOS ONE* 13.5 (2018), pp. 1–14 (cit. on p. 9).
- [53] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551 (cit. on p. 11).
- [54] S. C. Lee, M. Seibold, P. Fürnstahl, M. Farshad, and N. Navab. "Pivot calibration concept for sensor attached mobile c-arms". In: *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 11315. International Society for Optics and Photonics. SPIE, 2020, p. 1131503 (cit. on p. 2).
- [55] R. B. Lindsay. "The Wheel of Acoustics". In: *Journal of the Acoustical Society of America* 36 (1964), p. 2242 (cit. on pp. 17, 18).
- [56] M. P. E. Littré. *Oeuvres complètes d'Hippocrates*. Vol. 10. Baillière, 1861 (cit. on p. 5).
- [57] Z. Liu, J. Huang, and Y. Wang. "Classification TV programs based on audio information using hidden Markov model". In: *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No.98EX175)*. 1998, pp. 27–32 (cit. on p. 25).
- [58] L. Lu, H. Jiang, and H. Zhang. "A Robust Audio Classification and Segmentation Method". In: *Proceedings of the Ninth ACM International Conference on Multimedia*. 2001, pp. 203–211 (cit. on p. 25).
- [59] D. A. Lyon. "The Discrete Fourier Transform, Part 4: Spectral Leakage". In: *Journal of Object Technology* 8.7 (2009), pp. 23–34 (cit. on p. 22).

- [60] Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang. “LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm”. In: *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 2019, pp. 1–4 (cit. on pp. 8, 13).
- [61] A. Madhu and S. Kumaraswamy. “Data Augmentation Using Generative Adversarial Network for Environmental Sound Classification”. In: *2019 27th European Signal Processing Conference (EUSIPCO)*. 2019 (cit. on p. 63).
- [62] M. Margaryan, M. Seibold, I. Joshi, M. Farshad, P. Fürnstahl, and N. Navab. “Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data”. In: *arXiv preprint, arXiv:2211.02874* () (cit. on p. 1).
- [63] A. Marshall and S. Boussakta. “Signal analysis of medical acoustic sounds with applications to chest medicine”. In: *Journal of the Franklin Institute* 344.3 (2007), pp. 230–242 (cit. on p. 8).
- [64] S. Maslen. “Sensory Work of Diagnosis: A Crisis of Legitimacy”. In: *The Senses and Society* 11.2 (2016), pp. 158–176 (cit. on p. 7).
- [65] A. Massalimova, M. Timmermans, H. Esfandiari, F. Carrillo, C. J. Laux, M. Farshad, K. Denis, and P. Fürnstahl. “Intraoperative tissue classification methods in orthopedic and neurological surgeries: A systematic review”. In: *Frontiers in Surgery* 9 (2022) (cit. on p. 7).
- [66] S. Matinfar, T. Hermann, M. Seibold, P. Fürnstahl, M. Farshad, and N. Navab. “Sonification for Process Monitoring in Highly Sensitive Surgical Tasks”. In: *Nordic Sound and Music Computing Conference 2019 (Nordic SMC 2019)*. 2019 (cit. on p. 2).
- [67] S. Matinfar, M. Salehi, D. Suter, M. Seibold, N. Navab, S. Dehghani, F. Wanivenhaus, P. Fürnstahl, M. Farshad, and N. Navab. “Sonification as a Reliable Alternative to Conventional Visual Surgical Navigation”. In: *arXiv preprint, arXiv:2206.15291* () (cit. on p. 1).
- [68] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. 2020, pp. 405–421 (cit. on p. 12).
- [69] I. Miliaresi, K. Poutos, and A. Pikrakis. “Combining acoustic features and medical data in deep learning networks for voice pathology classification”. In: *2020 28th European Signal Processing Conference (EUSIPCO)*. 2021, pp. 1190–1194 (cit. on pp. 8, 13).
- [70] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny. “Deep Belief Networks using discriminative features for phone recognition”. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, pp. 5060–5063 (cit. on p. 11).
- [71] I. Morohashi, H. Iwase, A. Kanda, T. Sato, Y. Homma, A. Mogami, O. Obayashi, and K. Kaneko. “Acoustic pattern evaluation during cementless hip arthroplasty surgery may be a new method for predicting complications”. In: *SICOT-J* 3 (13 2017) (cit. on p. 10).
- [72] K. Nahen and A. Vogel. “Acoustic online monitoring of IR laser ablation of burnt skin”. In: *Laser-Tissue Interaction X: Photochemical, Photothermal, and Photomechanical*. Vol. 3601. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. 1999, pp. 392–397 (cit. on p. 11).
- [73] S. Nalband, A. Sundar, A. A. Prince, and A. Agarwal. “Feature selection and classification methodology for the detection of knee-joint disorders”. In: *Computer Methods and Programs in Biomedicine* 127 (2016), pp. 94–104 (cit. on p. 9).
- [74] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan. “Speech Recognition Using Deep Neural Networks: A Systematic Review”. In: *IEEE Access* 7 (2019), pp. 19143–19165 (cit. on p. 12).
- [75] J. Navarro, E. Vidaña-Vila, R. M. Alsina-Pagès, and M. Hervás. “Real-Time Distributed Architecture for Remote Acoustic Elderly Monitoring in Residential-Scale Ambient Assisted Living Scenarios”. In: *Sensors* 18.8 (2018) (cit. on p. 9).

- [76] S. Oberst, J. Baetz, G. Campbell, F. Lampe, J. C. S. Lai, N. Hoffmann, and M. Morlock. “Vibro-acoustic and nonlinear analysis of cadavric femoral bone impaction in cavity preparations”. In: *International Journal of Mechanical Sciences* 144 (2018), pp. 739–745 (cit. on p. 10).
- [77] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. “WaveNet: A Generative Model for Raw Audio”. In: *Arxiv*. 2016 (cit. on pp. 13, 63).
- [78] D. Ostler, M. Seibold, J. Fuchtmann, N. Sann, H. Feussner, D. Wilhelm, and N. Navab. “Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions”. In: *International Journal of Computer Assisted Radiology and Surgery* (5 2020) (cit. on pp. 2, 11, 13, 89).
- [79] J. H. Palep. “Robotic assisted minimally invasive surgery”. In: *Journal of Minimal Access Surgery* 5 (1 2009), pp. 1–7 (cit. on p. 11).
- [80] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. “Librispeech: An ASR corpus based on public domain audio books”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210 (cit. on p. 63).
- [81] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”. In: *Interspeech 2019* (2019) (cit. on p. 63).
- [82] K. Pauley, R. Flin, S. Yule, and G. Youngson. “Surgeons’ intraoperative decision making and risk management”. In: *The American Journal of Surgery* 202.4 (2011), pp. 375–381 (cit. on p. 7).
- [83] A. D. Pierce. *Acoustics*. Vol. 3. Springer Nature Switzerland, 2019 (cit. on p. 17).
- [84] “pix2xray: converting RGB images into X-rays using generative adversarial networks”. In: *International Journal of Computer Assisted Radiology and Surgery* 15 (2020), pp. 973–980 (cit. on p. 2).
- [85] B. M. Pohl, J. O. Jungmann, O. Christ, and U. G. Hofmann. “Automated drill-stop by SVM classified audible signals”. In: *34th Annual International Conference of the IEEE EMBS*. San Diego, California, USA, 2012 (cit. on p. 10).
- [86] M. Praamsma, H. Carnahan, D. Backstein, C. J. Veillette, D. Gonzalez, and A. Dubrowski. “Drilling sounds are used by surgeons and intermediate residents, but not novice orthopedic trainees, to guide drilling motions”. In: *Canadian Journal of Surgery* 51.6 (2008), pp. 442–446 (cit. on p. 9).
- [87] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-y. Chang, and T. Sainath. “Deep Learning for Audio Signal Processing”. In: *IEEE Journal on Selected Topics in Signal Processing* 14 (8 2019), pp. 206–219 (cit. on pp. 12, 23, 25, 26, 63).
- [88] A. Qurthobi, R. Maskeliūnas, and R. Damaševičius. “Detection of Mechanical Failures in Industrial Machines Using Overlapping Acoustic Anomalies: A Systematic Literature Review”. In: *Sensors* 22 (10 2022) (cit. on p. 13).
- [89] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. In: *arXiv e-prints*, arXiv:2204.06125 (2022) (cit. on p. 12).
- [90] G. W. Rodgers, J. L. Young, A. V. Fields, R. Z. Shearer, T. B. F. Woodfield, G. J. Hooper, and J. G. Chase. “Acoustic Emission Monitoring of Total Hip Arthroplasty Implants”. In: *IFAC Proceedings Volumes* 47.3 (2014). 19th IFAC World Congress, pp. 4796–4800 (cit. on p. 8).
- [91] D. Rumelhart, G. Hinton, and R. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323 (1986), pp. 533–536 (cit. on p. 11).
- [92] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252 (cit. on pp. 11, 94).

- [93] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: *arXiv e-prints*, arXiv:2205.11487 (2022) (cit. on p. 12).
- [94] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz. “Smartphone-based real-time classification of noise signals using subband features and random forest classifier”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 2204–2208 (cit. on p. 25).
- [95] J. Salamon, C. Jacoby, and J. P. Bello. “A Dataset and Taxonomy for Urban Sound Research”. In: *22nd ACM International Conference on Multimedia (ACM-MM’14)*. Orlando, FL, USA, 2014, pp. 1041–1044 (cit. on pp. 63, 94).
- [96] A. Schaufler, T. Sühn, N. Esmaeili, A. Boese, C. Wex, R. Croner, M. Friebe, and A. Illanes. “Automatic differentiation between Veress needle events in laparoscopic access using proximally attached audio signal characterization”. In: *Current Directions in Biomedical Engineering* 5.1 (2019), pp. 369–371 (cit. on p. 11).
- [97] R. Schwarzkopf, F. J. Kummer, and W. L. Jaffe. “Acoustic emission studies of posterior stabilized and cruciate retaining knee arthroplasties”. In: *Journal of Knee Surgery* 24 (3 2011), pp. 185–189 (cit. on p. 8).
- [98] M. Seibold, A. Hoch, M. Farshad, N. Navab, and P. Fürnstahl. “Conditional Generative Data Augmentation for Clinical Audio Datasets”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022, pp. 345–354 (cit. on p. 1).
- [99] M. Seibold, A. Hoch, D. Suter, M. Farshad, P. O. Zingg, N. Navab, and P. Fürnstahl. “Acoustic-Based Spatio-Temporal Learning for Press-Fit Evaluation of Femoral Stem Implants”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021, pp. 447–456 (cit. on p. 2).
- [100] M. Seibold, S. Maurer, A. Hoch, P. Zingg, M. Farshad, N. Navab, and P. Fürnstahl. “Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery”. In: *Scientific Reports* 11 (3993 2021) (cit. on p. 2).
- [101] M. Seibold, B. Sigrist, T. Götschi, J. Widmer, S. Hodel, M. Farshad, N. Navab, P. Fürnstahl, and C. J. Laux. “A new sensing paradigm for the vibroacoustic detection of pedicle screw loosening”. In: *arXiv preprint, arXiv:2210.16170* () (cit. on p. 1).
- [102] C. Shannon. “Communication in the Presence of Noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21 (cit. on p. 19).
- [103] M. Y. Shao, T. Vagg, M. Seibold, and M. Doughty. “Towards a Low-Cost Monitor-Based Augmented Reality Training Platform for At-Home Ultrasound Skill Development”. In: *Journal of Imaging* 8 (11 2022), p. 305 (cit. on p. 1).
- [104] M. Sharma and U. R. Acharya. “Analysis of knee-joint vibroarthrographic signals using bandwidth-duration localized three-channel filter bank”. In: *Computers & Electrical Engineering* 72 (2018), pp. 191–202 (cit. on p. 9).
- [105] N. P. Shine, P. G. O’Sullivan, J. Connell, P. Rulikowski, and J. Barrett. “Digital Spectral Analysis of the Drill-Bone Acoustic Interface During Temporal Bone Dissection”. In: *Otology & Neurotology* 27 (5 2006), pp. 728–733 (cit. on p. 10).
- [106] T. Song, M. Sommersperger, T. A. Baran, M. Seibold, and N. Navab. “HAPPY: Hip Arthroscopy Portal Placement Using Augmented Reality”. In: *Journal of Imaging* 8.11 (2022) (cit. on p. 1).
- [107] S. S. Stevens, J. Volkman, and E. B. Newman. “A Scale for the Measurement of the Psychological Magnitude Pitch”. In: *The Journal of the Acoustical Society of America* 8 (3 1937) (cit. on p. 21).

- [108] T. Sühn, A. Pandey, M. Friebe, A. Illanes, A. Boese, and C. Lohman. “Acoustic sensing of tissue-tool interactions – potential applications in arthroscopic surgery”. In: *Current Directions in Biomedical Engineering* 6.3 (2020), pp. 595–598 (cit. on p. 11).
- [109] Y. Sun, H. Jin, Y. Hu, P. Zhang, and J. Zhang. “State recognition of bone drilling with audio signal in Robotic Orthopedics Surgery System”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Chicago, Illinois, USA, 2014, pp. 3503–3508 (cit. on p. 10).
- [110] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari. “Phase reconstruction from amplitude spectrograms based on directional-statistics deep neural networks”. In: *Signal Processing* 169 (2020), p. 107368 (cit. on p. 22).
- [111] A. Tijou, G. Rosi, R. Vayron, H. A. Lomami, P. Hernigou, C.-H. Flouzat-Lachaniette, and G. Haiat. “Monitoring cementless femoral stem insertion by impact analyses: An in vitro study”. In: *Journal of the Mechanical Behavior of Biomedical Materials* 88 (2018), pp. 102–108 (cit. on p. 10).
- [112] Y. Torun and O. Pazarci. “Parametric Power Spectral Density Estimation-Based Breakthrough Detection for Orthopedic Bone Drilling with Acoustic Emission Signal Analysis”. In: *Acoustics Australia* (2020) (cit. on p. 10).
- [113] T. Tran, N. T. Pham, and J. Lundgren. “A deep learning approach for detecting drill bit failures from a small sound dataset”. In: *Scientific Reports* 12 (2022) (cit. on p. 94).
- [114] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. 2017 (cit. on p. 11).
- [115] C. Vununu, K.-S. Moon, S.-H. Lee, and K.-R. Kwon. “A Deep Feature Learning Method for Drill Bits Monitoring Using the Spectral Analysis of the Acoustic Signals”. In: *Sensors* 18.8 (2018) (cit. on p. 13).
- [116] O. Weede, F. Dittrich, H. Wörn, B. Jensen, A. Knoll, D. Wilhelm, M. Kranzfelder, A. Schneider, and H. Feussner. “Workflow analysis and surgical phase recognition in minimally invasive surgery”. In: *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2012, pp. 1080–1074 (cit. on p. 11).
- [117] J. C. J. Wei, W. H. A. Crezee, H. Jongeneel, T. S. A. De Haas, W. L. A. Kool, B. J. Blaauw, J. Dankelman, and T. Horeman. “Using Acoustic Vibrations as a Method for Implant Insertion Assessment in Total Hip Arthroplasty”. In: *Sensors* 22 (4 2022) (cit. on p. 10).
- [118] S. Wei, S. Zou, F. Liao, and weimin lang. “A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification”. In: *Journal of Physics: Conference Series* 1453.1 (2020), p. 012085 (cit. on p. 63).
- [119] J. M.-T. Wu, M.-H. Tsai, Y. Z. Huang, S. H. Islam, M. M. Hassan, A. Alelaiwi, and G. Fortino. “Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model”. In: *Applied Soft Computing* 78 (2019), pp. 29–40 (cit. on p. 8).
- [120] T.-c. I. Yang and H. Hsieh. “Classification of acoustic physiological signals based on deep learning neural networks with augmented features”. In: *2016 Computing in Cardiology Conference (CinC)*. 2016, pp. 569–572 (cit. on pp. 8, 13).
- [121] Z. Ying, L. Shu, and N. Sugita. “Bone Milling: On Monitoring Cutting State and Force Using Sound Signals”. In: *Chinese Journal of Mechanical Engineering* 25 (2022) (cit. on p. 10).
- [122] V. Zakeri and A. J. Hodgson. “Automatic Identification of Hard and Soft Bone Tissues by Analyzing Drilling Sounds”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (2 2019), pp. 404–414 (cit. on p. 10).

- [123] V. Zakeri and A. J. Hodgson. “Classifying hard and soft bone tissues using drilling sounds”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Seogwipo, South Korea, 2017 (cit. on pp. 10, 25).
- [124] Y. Zhang, B. Li, H. Fang, and Q. Meng. “Spectrogram Transformers for Audio Classification”. In: *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*. 2022, pp. 1–6 (cit. on p. 94).

List of Figures

1.1	The Laennec stethoscope. The examining physicians would place the wooden stethoscope on the patient’s chest and their ear on the other end of the tubular structure to screen heart and lung sounds for abnormalities.	6
2.1	The wheel of acoustics as defined by R. B. Lindsay, Adapted from [55].	18
2.2	a) A simple sine wave which illustrates the properties amplitude (highest derivation from central position) and wavelength (one full sine cycle). b) A real world waveform is composed of many different frequency components. The waveform is taken from the dataset created within the contribution presented in section 3.2.1 containing structure-borne drilling sound and a breakthrough event. . . .	19
2.3	The relationship between signal and frequency domain. The signal in the time domain is composed of two frequency components which are visible in the frequency spectrum as two peaks. The transformation from time to frequency domain is computed using the Fast Fourier Transform.	21
2.4	Spectrogram visualization of an audio sample taken from the dataset of the work presented in section 3.2.2, all spectrograms are computed from the same source waveform. The visualizations illustrate, from left to right, an amplitude spectrogram with linear frequency scaling as computed using equation 2.8, a power STFT spectrogram as described in equation 2.9 and linear frequency scaling, the same power STFT spectrogram with a logarithmic frequency scaling, and rightmost, a Mel power spectrogram.	22

List of Tables

1.1	The three types of machine learning including definitions, typical problems and example applications.	12
-----	---	----

